PUBLICATIONS OF GODDARD SPACE FLIGHT CENTER 1963

Volume I
SPACE SCIENCES

PUBLICATIONS OF GODDARD SPACE FLIGHT CENTER 1963

Volume I
SPACE SCIENCES

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C., 20402 - Price $9.75

FOREWORD

This publication is a collection of articles, papers, talks, and reports generated by the scientific and engineering staff of Goddard Space Flight Center in the year 1963. Many of these articles were originally published in scientific or engineering Journals or as official NASA technical publications, while others are documents of a more informal nature. All are reprinted here as nearly verbatim as typography and format will permit.

The articles are grouped into broad subject categories, but no detailed subdivision has been made. Within each category, the articles are arranged alphabetically by author. An overall author index is given in the back of the volume. A special Acknowledgments section is also included, which lists the various journals and other non-NASA publications from which particular articles have been taken.

The years 1963 and 1964 are being published as whole-year issues, and the resulting size dictates the use of two volumes; the first volume is titled Space Sciences, the second Space Technology. It is anticipated, however, that future issues will be quarterly single volumes.

CONTENTS

III. SOLAR PHYSICS

IV. IONOSPHERE AND RADIO PHYSICS

V. FIELDS AND PARTICLES

VI. PLANETOLOGY

VII. PLANETARY ATMOSPHERES

VIII. GENERAL

Not these

I. ASTRONOMY AND ASTROPHYSICS

INTERPLANETARY DUST

W. M. ALEXANDER

Goddard Space Flight Center

INTRODUCTION

Knowledge concerning the origin, composition, and dynamic properties of the cloud of dust particles surrounding the sun is fundamental to any extensive study of the solar system. Prior to the advent of rockets, satellites, and space probes, information concerning the physical and dynamic characteristics of dust particles in space came from a variety of ground-based experiments. Meteor observations, using visual, photographic, and radar techniques, produced measurements of the mass, velocity, density, and composition of individual meteoroids colliding with the earth's atmosphere. These studies were limited to particle masses greater than 10^{-4} gm until Gallagher and Eshleman (1960) and Eshleman (1962) reported radar observations of particles with limiting masses of approximately 10^{-6} gm. For particles with masses less than 10^{-6} gm, photometric observations of the zodiacal light, solar corona, and sky radiance have represented the major optical techniques used to obtain measurements of the sizes and dynamics of dust particles near the earth and in deep space. Composition, age, and gross accretion rates of meteoroidal and dust particle material have been obtained from analysis of collections of remnants of these particles surviving passage through the earth's atmosphere. These collection studies include dust particles with masses as low as 10^{-10} gm. The direct study of the physical and dynamic properties of dust particles in space with masses between 10^{-13} gm to 10^{-6} gm started with the measurements from dust particle experiments on various rockets, satellites, and space probes. Measurements obtained from ground-based and space experiments concerning dust particles with masses less than 10^{-6} gm and analyses of these results are summarized in the following sections.

SUMMARY OF DIRECT MEASUREMENTS

Six different types of sensors have been used to measure dust particle parameters including momentum, kinetic energy, and penetration and fracture properties associated with hypervelocity impact. An acoustical transducer, mechanically coupled to a metallic impact plate, has been the sensor used in most of the dust particle space experiments. Low velocity laboratory calibrations have shown this detector to be sensitive to the momentum of the impacting particles. A sensor which detects a light flash when impacted by a hypervelocity particle has also been used and provided a measurement associated with the kinetic energy of the detected dust particle. The results of these measurements are summarized in table 1.

The measurements from the first dust particle satellite experiments, Dubin (1960) and LaGow and Alexander (1960), showed that the average flux of dust particles near the earth was higher than expected from extrapolations of meteor observations. Also, Alexander et al. (1961), reported a marked increase in the dust particle flux rate during the period of the annual Leonid shower, figure 1. Dubin (1960) reported a "sporadic shower" measurement obtained from the Explorer I experiment and Dubin et al. (1962), summarized all the available evidence from direct measurements of cosmic dust showers.

McCracken et al. (1961) reported the initial results from the Explorer VIII experiment which used an acoustical transducer. An average cumulative mass distribution curve for dust particles

TABLE 1.—*Direct Measurements Obtained with Acoustical Transducer and Light Flash Systems on U.S. Satellites and Space Probes*

Spacecraft and sensor	Sensitivity dyne-sec$\times10^{-3}$ (acoustical)	Exposure (m^2-sec)	Number of particles	Omnidirectional influx rate $\left(\frac{\text{particles}}{m^2\text{-sec}}\right)$
Explorer I[a] (Acoustical)	>0.5	1.8×10^4	145	1.7×10^{-2}
Vanguard III[b] (Acoustical)	>10	2.8×10^6	~3500	2.0×10^{-3}
Pioneer I[c] (Acoustical)	>0.15	4.2×10^3	17	4.0×10^{-3}
Explorer VIII[d] (Acoustical)	2.5–24	2.4×10^5	~3650	3.0×10^{-2}
	25–250		~75	6.2×10^{-4}
	>250		1 or 2	$\sim1.0\times10^{-5}$
Ranger I[e] (Acoustical)	>0.030	8.8	64	4.0×10
Midas II[f] (Acoustical)	>0.3	2.7×10^3	67	5.0×10^{-1}
Samos II[g]	$\gtrsim$0.3	----------	----------	6.8×10^{-1}
	grams (light flash)			
Explorer VIII[h] (Light flash)	>1 (prelim.)	4.3	110	2.0×10^2
Ranger I[i]	>1 (prelim.)	8.8	179	1.1×10^2

Initial reports:

(a) Dubin (1960); (b) LaGow and Alexander (1960); (c) Dubin (1960; (d) McCracken et al. (1961); (e) Alexander and Berg (1962); (f) Soberman and Della Luca (1961); (g) Soberman and Della Luca (1961); (h) Alexander et al. (1962); (i) Alexander and Berg (1962).

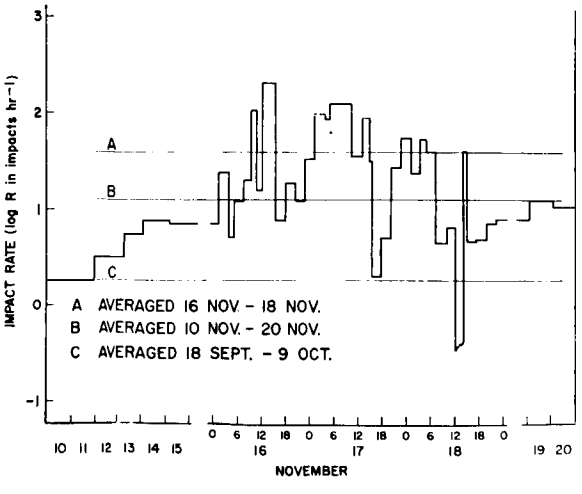


FIGURE 1.—Impact rates during the November 1959 interplanetary dust particle event.

near the earth was obtained from this measurement. This curve was markedly different from that expected from extrapolations of results from meteor observations. McCracken (1962) compared the available results of all the previous experiments having acoustical sensors to the distribution curve obtained from the Explorer VIII experiment and found excellent agreement.

All the available direct measurements obtained with microphone and light flash sensors were reviewed by Alexander et al. (1962), and an average cumulative mass distribution curve for dust particles near the earth with masses between 10^{-13} gm and 10^{-6} gm was presented. This curve is shown in figure 2.

Hemenway and Soberman (1961, 1962) reported the results of analyses of dust particles

collected from a recoverable sounding rocket experiment. These results are also shown in figure 2 and compared to the composite distribution curve.

Two measurements of dust particle flux from space probes have been reported. Dubin (1960) showed that the average flux in cis-lunar space was approximately 10^{-2} times that near the earth. The initial results from an experiment in interplanetary space (Alexander, 1962), showed a dust particle flux 10^{-4} times that near the earth.

Results of experiments using sensors detecting damage occurring during an impact were reported by LaGow et al. (1961), and Davison et al. (1961), and compared to the data from all other direct measurement experiments by Alexander et al. (1962).

GROUND BASED MEASUREMENTS

Collections and analyses of remnants of meteoroidal and sut particle material from ocean depths, mountain tops, and the Antarctic Ice Cap were reported by Petterson (1960), Crozier (1960, 1961, 1962) and Thiel and Schmidt (1961). The measurements of annular accretion rates as a function of particle size (5 to 15 μ diameter) from collectors placed at isolated high altitude positions in New Mexico and analysis of Antarctic Ice Cores are in excellent agreement. The sediment cores from ocean depths indicate an annular deposit which is one to two orders of magnitude less than the other measurements. Crozier (1962) analyzed the results of the various collection studies and the dust particle distribution reported by McCracken et al. (1961), and found an agreement within an order of magnitude for particles in a mass range of 10^{-9} gm to 10^{-7} gm.

The results of an extensive set of measurements from observations of the zodiacal light and the solar corona were reported by Ingham and Blackwell (1961). Dubin and McCracken (1962) compared the results of these measurements and some of the earlier studies of the zodiacal light and solar F corona with the results of direct measurements from various dust particle experiments on satellites. This comparison of the spatial densities derived from direct measurements and zodiacal light studies indicated that the

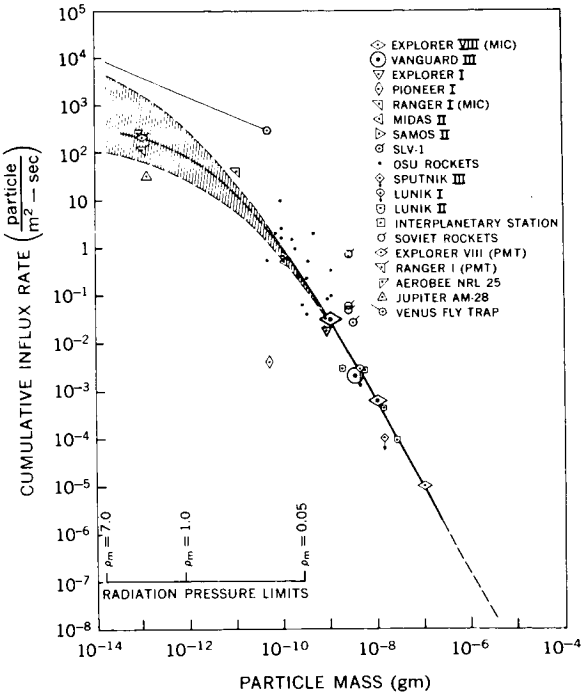


FIGURE 2.—An average cumulative mass distribution curve for the vicinity of earth derived from all the available direct measurements obtained with microphone and photomultiplier systems. The influx rates are omnidirectional values. A dust particle of mass density ρ_m is expelled from the solar system by radiation pressure if its mass is less than that shown on the lower abscissa opposite the particular value of ρ_m.

spatial density of dust particles near the earth was approximately 10^3 times greater than in interplanetary space.

Newkirk and Eddy (1862) studied the influx of meteor particles in the upper atmosphere of the earth (80,000 feet) from balloon coronagraph observations. They showed an influx rate which was less than the measurements from satellite experiments but in excess of expected influx rates from extrapolations of ground based meteor observations.

ADDITIONAL ANALYSES

The results of direct measurements concerning the flux of dust particles near the earth has given added emphasis to considerations of a concentration of dust particles near the earth. Whipple (1960, 1961) compared satellite dust particle measurements and zodiacal light and solar corona

photographic observations and concluded that there was a concentration of dust near the earth. Hibbs (1961) presented evidence for a dust cloud near the earth from a statistical evaluation of the altitude dependence of dust particle flux obtained from the Explorer I experiment. These results were further discussed by Dubin (1961) and Hibbs (1961) concerning the statistical significance of the data available from this experiment.

A group of letters by Bear (1961), Fremlin (1961) and Whipple (1961) discussed the existing evidences of such a dust cloud about the earth and possible theoretical mechanisms for creating such a cloud. In particular, the mechanism of gravitational attraction was explored. Singer (1961) carefully examined the role of gravitational accretion as a mechanism for creating a dust cloud and concluded that this factor would enhance the dust particle flux near the earth as compared to deep space by no more than a factor of ten.

Whipple (1961) also suggested that the flux of dust particles near the earth could be ehnanced by lunar ejecta being trapped in the earth-moon gravitational field. Whipple (1962) reported on the studies of meteoritic erosion in space as one of the contributing mechanisms to the spatial density of dust particles in interplanetary space and showed that these studies were in general agreement with representative zodiacal light and solar corona photographic observations. He concluded, as did Dubin and McCracken (1962), that these measurements indicated an interplanetary dust particle spatial density 10^{-3} times less than the satellite direct measurement results and this comparison showed good evidence for a dust cloud around the earth.

Carleton (1962) considered the relation of the atmospheric dust measurements of Volz and Goody (1962) to the problem of meteoric influx. His analysis showed that the dust particle flux near the earth, as reported by McCracken et al. (1961), was a more than adequate source for the measured atmospheric dust. In fact, less than 1% of the satellite measured flux was needed to produce the required influx of dust into the atmosphere. Thus, he concludes that there is a reservoir of orbital dust forming a cloud around the earth.

The strongest additional evidence of a concentration of dust particles near the earth was the direct measurement of dust particle flux in interplanetary space from the Mariner II experiment which showed a flux approximately 10^{-4} times that near the earth. The satellite results have produced a measurement of the mass distribution of dust particles in this cloud. An extensive measurement of the dynamics of these dust particles has not been made and the mechanism by which this cloud is created is not completely understood.

REFERENCES

Alexander, W. M., McCracken, C. W., and LaGow, H. E., Interplanetary dust particles of micron-size probably associated with the Leonid meteor stream, *J. Geophys. Research*, **66**, 3970–3973, 1961.

———, and Berg, O. E., Microparticle hypervelocity impacts from Ranger I, published in *Proceedings of Fifth Hypervelocity Symposium*, **1962**.

———, McCracken, C. W., Secretan, L., and Berg, O. E., Review of direct measurements of interplanetary dust from satellites and probes, *Space Research III*, North-Holland Publishing Co., Amsterdam, Proceedings of COSPAR, 1962 (to be published). Also available NASA TN D-1669, 1962.

———, Preliminary results of cosmic dust experiment on Mariner II, Science, 138–3545, 1098–1099, Dec **7**, 1962.

Beard, D. B., The dust cloud about the earth, *Nature*, **191**, 32–33, 1961.

Carleton, N. P., The relation of the recent atmospheric dust measurements of Volz and Goody to the problem of meteoric influx, *J. Atmospheric Sciences*, **19**, 424–426, 1962.

Crozier, W. D., Black, magnetic spherules in sediments, *J. Geophys. Research*, **65**, 2971–2977, 1960.

———, Micrometeorite measurements: satellite and ground level data compared, *J. Geophys. Research*, **66**, 2793–2795, 1961.

———, Five years of continuous collection of black, magnetic spherules from the atmosphere, *J. Geophys. Research*, **67**, 2543–2548, 1962.

Davison, E. H., and Winslow, P. C., Jr., Space debris hazard evaluation, NASA TN D-1105, 1961.

Dubin, M., IGY Micrometeorite measurements, *Space Research I*, North Holland Publishing Co., Amsterdam, 1042–1058, 1960.

———, Meteoritic dust measured from Explorer I, *Planetary and Space Science*, **2**, 121–129, 1960.

———, Remarks on the article by A. R. Hibbs, "The distribution of micrometeorites near the earth," *J. Geophys. Research*, **66**, 2592–2594, 1961.

———, ALEXANDER, W. M., and BERG, O. E., Direct measurements of cosmic dust showers, presented at the *International Symposium on the Astronomy and Physics of Meteors*, proceedings published in *Smithsonian Contributions to Astrophysics*, 1962.

———, and MCCRACKEN, C. W., Measurements of distributions of interplanetary dust, *Astronomical Journal*, **67**, 248–256, 1962.

ESHLEMAN, V. R., Radar rate measurements on very small meteors, *Astronomical Journal*, **67**, 242–248, 1962.

FREMLIN, J. H., The dust cloud about the earth, *Nature*, **191**, 31–32, 1961.

GALLAGHER, P. B., and ESHLEMAN, V. R., Sporadic shower properties of very small meteors, *J. Geophys. Research*, **65**, 1846–1847, 1960.

HEMENWAY, C. L., and SOBERMAN, R. K., Studies of micrometeorites obtained from a recoverable sounding rocket, *Astronomical Journal*, **67**, 256–266, 1962.

HIBBS, A. R., The distribution of micrometeorites near the earth, *J. Geophys. Research*, **66**, 371–377, 1961.

———, Author's reply to the preceding discussion on the article, "The Distribution of Micrometeorites Near the Earth," *J. Geophys. Research*, **66**, 2595–2596, 1961.

INGHAM, M. F., Observations of the zodiacal light from a very high altitude station IV (also the three preceding papers, p. 113–155, by D. E. Blackwell and M. F. Ingham), *Mon. Not. R.A.S.*, **122**, 157–176, 1961.

LAGOW, H. E., and ALEXANDER, W. M. Recent direct measurements of cosmic dust in the vicinity of the earth using satellites, *Space Research I*, North Holland Publishing Co., Amsterdam, 1033–1041, 1960.

———, and SECRETAN, L., The micrometeorite penetration experiment on Explorer VII, NASA TN D–608, 1961.

MCCRACKEN, C. W., ALEXANDER, W. M., and DUBIN, M., Direct measurements of interplanetary dust particles in the vicinity of earth, *Nature*, **192**, 441–442, 1961.

———, and ———, The distribution of small interplanetary dust particles in the vicinity of earth, presented at the *International Symposium on the Astronomy and Physics of Meteors*, proceedings published in *Smithsonian Contributions to Astrophysics*, 1962.

NEWKIRK, GORDON, Jr., and EDDY, J. A., Influx of meteor particles in the upper atmosphere of the earth as determined from stratospheric coronograph observations, *Space Research III*, North Holland Publishing Co., Amsterdam, Proceedings of COSPAR, 1962 (to be published).

PETTERSSON, H., Cosmic spherules and meteoritic dust, *Sci. American*, **202**, 123–132, 1960.

SINGER, S. F., Interplanetary dust near the earth, *Nature*, **192**, 321–323, 1961.

SOBERMAN, R. S., and DELLA LUCCA, L., Micrometeorite measurements from the Midas II satellite, (1960–61) G.R.D. Res. Note No. 72, 1961.

———, and HEMENWAY, C. L., et al., Micrometeorite collection from a recoverable sounding rocket, G.R.D. Research Note, No. 72, 1961.

THIEL, E., and SCHMIDT, R. A., Spherules from the antarctic ice cap, *J. Geophys. Research*, **66**, 307–310, 1961.

VOLZ, F. E., and GOODY, R. M., The intensity of the twilight and upper atmospheric dust. *J. Atmospheric Sciences*, **19**, 351–368, 1962.

WHIPPLE, FRED L., Particulare contents of space, *Proceedings of the Symposium on Medical and Biological Aspects of the Energies of Space at the Southwest Research Institute, San Antonio, Tex.*, Oct. 24, 1960.

———, The dust cloud about the earth, *Nature*, **189**, 127–128, 1961.

———, The dust cloud about the earth, *Nature*, **191**, 33–34, 1961.

———, Meteoritic erosion in space, *Space Research III*, North Holland Publishing Co., Amsterdam, Proceedings of COSPAR, 1962 (to be published).

INTERPLANETARY MAGNETIC FIELDS AS A CAUSE OF COMET TAILS

DAVID B. BEARD* AND M. PAUL NAKADA

Goddard Space Flight Center

It has been generally recognized recently that the historic explanation of comet tails in terms of solar radiation pressure by either light or particles is impossible. Solar light pressure has been found to be insufficient by several orders of magnitude. Coulomb collisions or acceleration coupled with charge transfer by protons in the solar wind have also been shown by Biermann and Treffitz,[1] in particular, to be grossly insufficient processes to account for comet tails streaming within a cylinder of small diameter and great length away from the Sun. It is the purpose of this communication to point out that the gases in cometary comas will be efficiently ionized by the solar wind of ionized hydrogen embedded in a magnetic field[2-4] and that an interplanetary magnetic field will couple the cometary gas to the solar wind.

The neutral gases evaporated from comet nuclei will be ionized both by thermal electrons[4] in the solar wind which have energies of several tens of electron volts and, as Heubner[5] has emphasized, by Alfvén's[6] process of ionization of a neutral gas by a fast moving plasma embedded in a magnetic field.

A comet head may be regarded as a densely ionized plasma essentially stationary in a high-pressure solar wind containing a low-pressure magnetic field moving with the wind. Recent satellite measurements[4] have established that the solar wind consists of plasma the protons of which have an isotropic thermal energy of a few electron volts and a kinetic energy due to their stream velocity of a kilovolt or so. The interplanetary magnetic field observed to be embedded in the plasma is about 5 γ (1 γ is 10^{-5} gauss). Thus the magnetic pressure, $H^2/8\pi$, is about equal to the thermal (isotropic) pressure of the wind in the moving co-ordinate frame of the wind.

The stream pressure of the solar wind, however, is the overwhelmingly dominant pressure in space, being two orders of magnitude larger than the free space magnetic pressure. Hence, the interplanetary magnetic field is compressed against any obstacle such as the stationary plasma provided by the comet coma or the stationary magnetic field provided by the geomagnetic dipole. Just as in the case of the example furnished by the geomagnetic dipole,[7] the interplanetary field is compressed against the obstacle offered by the comet coma so that the magnetic pressure increases to approximately equality with the stream pressure. If the interplanetary field is parallel to the solar wind velocity, the compressed field tails off parallel to the solar wind in a cylindrical shape for a large distance downstream from the comet coma confining the comet plasma within this volume. If the interplanetary field is perpendicular to the wind velocity, the compressed field lines on the solar side of the comet slip around the edge of the comet and then continue moving with the solar wind.[7] The comet plasma will become embedded in the moving field lines and will be carried radially away from the Sun, acquiring the velocity of the magnetic field as Heubner[5] first proposed.

Quite simply the magnetic field may be regarded as the mechanism by which the solar wind is efficiently coupled to the comet plasma which Coulomb collisions are not able to provide. Fluctuations in solar wind pressure, or the magnetohydrodynamic waves the fluctuations generate,

*National Academy of Sciences—National Research Council Research Associate with N.A.S.A. Permanent address: Physics Department, University of California, Davis.

will also create the accelerations in comet tails which first led Biermann[2] to deduce the presence of the solar wind.

REFERENCES

1. BIERMANN, L., and TREFFITZ, E., *Z. Astrophys.*, **49**, 111 (1960).
2. ———, *Mem. Soc. Roy. Sci. Liege*, Ser. **4, 13,** 291 (1953).
3. ———, *Nuovo Cimento*, Supp. **13,** 189 (1959).
4. NEUGEBAUER, M., and SNYDER, C., *Science*, **138,** 1095 (1962). COLEMAN, P. J., JR., DAVIS, L., JR., SMITH, E. J., and SONNETT, C. P., *Science*, **138,** 1099 (1962).
5. HEUBNER, W. F., *Revs. Mod. Phys.*, **33**, 498 (1961).
8. ALFVEN, H., *Rev. Mod. Phys.*, **32,** 710 (1960).
7. BEARD, D. B., *J. Geophys. Res.* (to be published).

FORMATION OF THE SOLAR NEBULA

A. G. W. CAMERON

Goddard Space Flight Center

J. Gaustad has shown that no pause is to be expected in the free-fall collapse of an interstellar cloud until the hydrogen and helium of the fragments become ionized. It follows that there cannot be significant loss of angular momentum from the fragments during the collapse. The size and mass distribution of the resulting nebular disks formed from the fragments are discussed. Following a discussion of Mestel, it is concluded that two forms of mass distribution are possible, each with dimensions comparable to those of the solar system. In one, with axial condensation, there is shear and further condensation to form a single star with accompanying planets seems likely. In the other, without axial condensation, there is no shear and binary star formation seems likely.

INTRODUCTION

In a recent paper (Cameron, 1962) I discussed a number of problems connected with the formation of the solar system. Among these was the question of star formation and how it can lead to the formation of a solar nebula from which the planets can be formed. The stages discussed were as follows:

1. The onset of gravitational instability in a cloud of interstellar gas.
2. The collapse of this cloud in free fall and its fragmentation into smaller sub-units.
3. The onset of a Helmholtz-Kelvin contraction stage in the fragments when the opacity becomes high enough to prevent the escape of the thermal energy generated by the contraction.
4. Gravitational instability and collapse of the fragments when the internal temperature rises high enough to start dissociating hydrogen molecules.
5. Further Helmholtz-Kelvin contraction after the internal temperature has risen sufficiently to doubly ionize helium throughout most of the interior.

J. Gaustad (1962) has recently calculated the opacity of matter at low temperatures and densities, and has arrived at the important conclusion that stage three above does not exist. By the time that the dissociation of hydrogen molecules commences, the center is at only a small number of optical depths. Hence the free-fall collapse is not checked.

This development necessitates some revisions in my discussion of the formation of the solar nebula, since it can no longer be assumed that the dissociation and ionization collapse starts from a state of hydrostatic equilibrium of a protostar. It is the purpose of this note to amend the previous discussion.

THE NEBULAR DISKS

During the collapse of the interstellar cloud, it appears that fragmentation will occur centered about density fluctuations large enough to individually satisfy the Jeans' instability criterion (Hunter, 1962). Hence the fragments are likely to have some sort of density maximum in the central regions, but this will be very mild compared to that in a structure in hydrostatic equilibrium. Hence for a discussion of the formation of the solar nebula it appears reasonable to take the fragment as a uniformly rotating, uniform sphere, and to find what kind of thin disk is formed when such a sphere collapses with local conservation of angular momentum.

This problem has been discussed in connection with the formation of galaxies by Mestel (1962). A brief adaptation of his discussion is given here.

It appears that the disk formed from the col-

lapse of the sphere can have either of two radically different density distributions. Consider a particle in such a disk. In one case the gravitational potential which governs the motion of the particle depends essentially only on the mass interior to the particle. This disk will be described as "axially condensed." In the other case the gravitational potential depends in a comparable way on the mass both interior to and exterior to the particle. This disk will be described as "uniform."

Consider first the uniform disk. It has a mass per unit area given by

$$\sigma=\sigma_0[1-(r/R)^2]^{1/2} \tag{1}$$

where σ_0 is a constant and R is the radius. A particularly significant property of this disk is its uniform rotation at the rate

$$\omega = [2\pi G\sigma_0/2R]^{1/2}, \tag{2}$$

where ω is the angular velocity and G is the gravitational constant. This disk can be formed from a uniform sphere with exact local conservation of angular momentum (MacMillan, 1930).

No precise analytical form is known for the axially condensed disk that can be formed from a sphere. However, Mestel points out that it probably approximately resembles the following form:

$$\sigma' r' = \text{constant} \tag{3}$$

$$\omega' r' = [2\pi G\sigma' r']^{1/2} = \text{constant}. \tag{4}$$

Here we have denoted by primed quantities the same physical variables used previously. This infinite disk has both the surface mass density and angular velocity varying inversely as the radial distance. The angular velocity is the same that it would be if all the mass interior to a given point were concentrated at the origin. Mestel points out that, at least near the center, the axially condensed disk can be derived with approximate local conservation of angular momentum using the transformation

$$r' = \tfrac{1}{2}\pi(r^2/R). \tag{5}$$

It seems sufficient for the purposes of this discussion to assume that the nebular disk of equation 3 can be obtained from collapse of a uniformly rotating uniform sphere. We shall cut off the mass distribution at some radius R'.

DIMENSIONS OF THE DISKS

One of the most significant aspects of Gaustad's work is to show that free-fall collapse of the interstellar material is not interrupted until either ionization of helium is completed, or the material flattens to form a disk. I have previously shown (Camerson, 1962) that the twist of the material in an interstellar cloud remains small during free-fall collapse provided it has an initial density similar to the densest observed clouds. This appears to be the optimum condition for the initiation of gravitational instability. Hence there appears to be no opportunity for any significant transmission of angular momentum by magnetic coupling to the surroundings. Thus we can assume that the material forming any final disk has the same angular momentum that it had as part of the interstellar cloud. This allows us to calculate the dimensions of the disks that may be formed.

Let us suppose that the original spherical volume had uniform density ρ, radius R_i, and angular velocity ω_i. If this material flattens into the uniform disk, then from the principles of conservation of mass and angular momentum we find

$$\sigma_0 = 2\pi^4 G^2 \rho^3 R_i/\omega_i^4 \tag{6}$$

$$R = \omega_i^2 R_i/\pi^2 G\rho \tag{7}$$

Note the great sensitivity of the dimensions of the disk to the initial density and angular velocity of the interstellar cloud. This leads us to expect that the radius of disks formed by the collapse of interstellar clouds may be subject to considerable variability. If we express the total mass M of the disk in solar units and the initial density of the cloud in terms of the number density of hydrogen atoms (assuming solar composition), then

$$\sigma_0 = 1.03\times10^{-4}\, M^{1/3} n_H^{8/3}\ \text{gm/cm}^2 \tag{8}$$

$$R = 2.08\times10^{5}\, M^{1/3} n_H^{-4/3}\ \text{a.u.} \tag{9}$$

In these expressions we have also assumed $\omega_i = 10^{-15}$ radians/second, corresponding to a rotation of the interstellar cloud once per orbital period in the galaxy.

For the axially condensed disk we find, in terms

of the analogous quantities of the uniform disk:

$$\sigma' r' = (25/36\pi)\sigma_0 R \tag{10}$$

$$R' = (12\pi/25)R. \tag{11}$$

It was previously argued (Cameron, 1962) that a reasonable choice for the initial density is $n_H = 10^3$ cm^{-3} and that M might lie in the range 1 to 10 solar masses. This gives us nebular disks with radii of a few tens of astronomical units, just of the same order as that of the solar system. Indeed, the mass distribution in the axially condensed disk then turns out to be rather similar to those found previously for the collapse of polytropic spheres (Cameron, 1962). Only the axially condensed disk has a central body in hydrostatic equilibrium, and its mass is very small compared to that of the sun.

FINAL STAR FORMATION

It is now natural to inquire how stars can be formed by further condensation from disks of the sort discussed here. Only qualitative suggestions can be made on this point at the present time.

If a uniform disk is formed, its most important characteristic feature is its uniform rotation. There is clearly no tendency in such a disk for a condensation toward the center. Indeed, the mass in any radial interval increases toward a peak at a radial distance of about 70% of the radius of the disk. Note that we could translate an element of mass around its orbit without changing its angular momentum.

This suggests that such a disk may be unstable against condensation into a binary star pair. This would arise from condensation around an orbit toward any common diameter of the disk, leading to the formation of two centers of mass concentration. If all the mass of the disk goes into a binary pair with equal mass division, then the radius of the resulting binary orbit is $(3\pi/25)R$, from conservation of angular momentum. With disks of the dimensions to be expected from the above discussion, this radius agrees well with the most frequently observed orbital radii of binary stars (Kuiper, 1951).

If an axially condensed disk is formed, then everywhere there is a shear within the gas. As was suggested previously (Cameron, 1962), any magnetic field trapped within the gas will be pulled out and strengthened by the shear. The energy of the field must be derived from the energy of differential rotation, and hence in order to maintain Kepler motion the majority of mass in the disk must move inwards, while a small portion of its moves outwards to conserve angular momentum. Thus there is a natural tendency for a central condensation to occur from such a disk, but it must be very difficult for a second major center of condensation to be formed against the tidal action of the central field. Hence the axially condensed disk seems likely to lead to the formation of single stars, and there is an opportunity for solid materials to condense out of the gas and form planets. However, the detailed processes of the planet formation remain to be determined.

REFERENCES

CAMERON, A. G. W., The Formation of the sun and the planets. *Icarus* **1**, 13, (1962).

GAUSTAD, J., Paper at Yale meeting of the American Astronomical Society, (1962).

HUNTER, C. (1962) (to be published).

KUIPER, G. P., *In* "Astrophysics" (J. A. Hynek, ed.). McGraw-Hill, New York, (1951).

MACMILLAN, W. D., *Theoretical Mechanics*, Cambridge University Press, Cambridge (1962).

MESTEL, L., Privately circulated manuscript and Institute for Advanced Study lecture notes (1962).

THE COLLAPSE PHASE OF EARLY SOLAR EVOLUTION

A. G. W. CAMERON

Goddard Space Flight Center

That phase of the contraction of a star in which H_2 molecules are being dissociated and H and He atoms are being ionized is examined, and it is found that a stellar configuration must be unstable against dynamical collapse during this phase. Quantitative calculations have been made with polytropes of indices 1.5 and 3. The gravitational instability sets in for stars of one solar mass at a radius of about 100 A.U. and ceases at about 1/3 A.U. If a star is rotating, having conserved angular momentum during its collapse, then it should flatten into a nebular disk without forming a central body in hydrostatic equilibrium. Some physical parameters of possible disks are given.

INTRODUCTION

The riddle of solar system formation is one of the oldest unsolved problems of scientific philosophy. Serious speculation on this topic dates back to Descartes. More than 300 years ago, he advanced some ideas concerning the condensation of the sun and planets from a chaotic mass of dust and gas; these speculations bear a remarkable resemblance to some of the ideas now accepted about the formation of the solar system. In the intervening three centuries, a great many schemes for the formation of the solar system have been put forward, some of them in remarkable mathematical detail; however, they are only as plausible as their assumed set of initial conditions, and the initial conditions vary enormously.

The basic difficulty with the development of such a theory is that we have very little concrete information upon which to base it. Until recent years, this information consisted only of a number of remarkable regularities observed within the solar system itself. There are basically two regularities involved. One is the tendency for alignment of the angular momentum vectors in the solar system. This means that: (1) the rotations of the sun, planets, and satellites tend to have the same equatorial planes, and (2) the motions of the satellites about the planets, and of the planets about the sun, tend also to lie in the same plane. The other regularity is that there seems to be a smoothly increasing spacing between the orbits of the planets. When we examine the nature of the regularities in orbital spacings, the latter, known as Bode's law, is often more impressive in its departures from strict regularity than in its adherence to it.

In more recent years we have obtained much additional information about the early history of the solar system for the study of meteorites. These objects are usually believed to be fragments of asteroidal bodies that have been destroyed by collisions. Thus, a study of the chemical and physical properties of meteorites can yield much information about conditions in the interiors of asteroidal bodies. This subject is extremely complicated, and the resulting interpretations are still very controversial.

One of the more exciting discoveries of recent years concerning the meteorites is the presence in some of them of anomalous isotopic compositions of certain elements. These anomalies take the form of excess abundances of isotopes which are decay daughters of certain radioactivities with half-lives of a few million years. It appears that these radioactivities, though now extinct, must have been present in the meteorite parent bodies when they were formed. From the abundances of these radioactivities (deduced from the anomalous abundance of the

daughter isotopes in the daughter elements), much can be inferred as to the time at which chemical isolation of the meteorite parent bodies occurred.

We may expect to obtain a great deal of additional information about the early history of the solar system when we begin systematic explorations of the moon. Because of: (1) The absence of erosion on its surface, and (2) its apparently minimal amount of volcanic activity, the moon has preserved a record of the solar system's early history that has long since been eradicated from the earth. Thus, within the next few years, it should become possible to extrapolate the history of the solar system backward in time with a great deal more assurance than can be done at present. It is doubtful, however, that we shall ever properly understand the processes involved in the formation of the solar system solely from studies of this kind. We must combine such findings with studies of star formation. Most investigators have now come to view the planets as having been formed by some sort of condensation process of a nebular disk of gas and dust that once encircled the sun. There is considerable controversy as to how this disk was formed, but many feel that it must have been a natural consequence of the sun's formation. Thus, it is desirable to study all phases of the star formation process.

STAR FORMATION

The study of star formation and its relation to the formation of the solar system has begun much more recently than has speculation about the formation of the solar system itself. Such studies were not very meaningful until recent years, when it became apparent that considerable gas and dust lying between the stars is available for condensation into new stellar systems; and that some stars have been formed in our galaxy within the last few million years. This has lead to much speculation about the conditions necessary for the formation of stars.

Other studies in recent years have shown that the galaxy is much older than had previously been thought. In particular, detailed studies of stellar evolution conducted with large electronic computers have given results that can be compared with the Hertzprung-Russell diagrams of old galactic and globular star clusters, and from which the ages of the clusters can be deduced. There is still much uncertainty about the resulting ages, but it appears that the galaxy is at least 1×10^{10}, and possibly as great as 2.5×10^{10} years old. For comparison, studies of the relative isotopic composition of lead in the earth and in the meteorites have shown that the solar system is some 4.5×10^{9} years old. Thus, the galaxy is very much older than the solar system. Since the galaxy was presumably composed entirely of gas in the beginning, and since that gas now constitutes only about 2 percent of the galactic mass, it seems evident that star formation must have gone on at a much greater rate in the early history of the galaxy than it does at present. It seems a reasonable conclusion from this that at the time the solar system was formed, the galaxy was not very different in appearance from that of today. Hence, we have some justification for believing that observations of star formation today are relevant to the conditions under which the sun was formed.

Nearly all studies of the star formation process have concluded that stars are probably not formed singly, but in associations or clusters resulting from the contraction of an interstellar cloud. The interstellar gas seems to be subject to great fluctuations in both its density and velocity distributions, the regions of greater density being called clouds. There appears to be a general concentration of the dense clouds toward the spiral arms of the galaxy (the positions where most of the newer stars are formed); there also appears to be a direct connection between the presence of dense clouds of gas and the formation of stars in the arms.

We can readily understand the tendency of new stars to form from dense interstellar clouds. Such clouds are rarely in equilibrium; they usually have a tendency to expand or contract. They will expand if their internal energy, in the form of heat and turbulent motion, is comparable to or greater than their self-gravitational potential energy. In such cases, the expansion will usually be limited by an external pressure, perhaps by the interstellar magnetic field or by surrounding gas at higher temperature. However, if the internal energy becomes small enough as compared to the

gravitational potential energy, the cloud tends to contract. Once such contraction is well under way, it is difficult to visualize its stopping short of star formation. As the density of the cloud increases, the efficiencies of its cooling processes also increase, and hence, its internal energy becomes very small as compared with its gravitational potential energy. This greatly reinforces the tendency toward contraction.

As the contraction of a cloud proceeds, the gravitational potential energy rapidly increases. Consequently, small parts of the cloud are able to contract individually, since the internal energy of the gas in such small regions becomes small compared with the gravitational potential energy of the region itself. Thus, the cloud can fragment into many different starts, ceasing when the ultimate fragment becomes opaque to its own radiation, so that the internal energy of its constituent gas can no longer decrease relative to its gravitational potential energy. Bodies have then formed with masses approximately equal to that of the sun. It is with the subsequent evolution of such bodies that we shall be mainly concerned in this discussion.

ANGULAR MOMENTUM

One of the principal questions with which we shall be concerned regarding the early evolution of a star, is the amount of angular momentum it contains. We have some idea as to the amount of angular momentum of the interstellar cloud before the contraction starts. The motions of the interstellar clouds seem to be constrained, at least to some extent, by the presence of an interstellar magnetic field. The ions of the interstellar gas, both those produced by the photoionization processes of starlight and those produced by the ionizing effects of cosmic ray particles, are bound in close spiral motions about the magnetic lines of force. The collisions of the neutral atoms in the interstellar gas with these ions are sufficient to cause all of the gas to stay closely associated with the magnetic lines of force. Thus motion along these lines is easily possible, while motion at right angles to them is very difficult. Consequently, the interstellar magnetic field must be trapped in the gas that condenses to form the stars. Before the condensation starts, the magnetic field may well constrain the motion of the interstellar cloud so that it rotates once per revolution about the center of the galaxy. This corresponds to an angular velocity of 10^{-15} rad/sec at the sun's distance from the center of the galaxy. Random motions of the interstellar clouds may well induce larger angular velocities, but this value is the minimum angular velocity that we might expect a cloud to possess.

When a cloud contracts and fragments to form a cluster or association of stars, a general conservation of angular momentum must prevail, and some of the cloud's initial angular momentum will go into the orbital motion of the star fragments about the center of mass. The fragments themselves will also rotate more and more rapidly as their contraction proceeds. The major question seems to be: What role, if any, is played by the trapped interstellar magnetic field in transmitting angular momentum from the contracting cloud to the stable surrounding interstellar medium?

A recent study by the present author indicates that the contracting cloud does not twist appreciably during its contraction and fragmentation stage, so that presumably very little angular momentum is transmitted to the interstellar medium at such a time. A more difficult question, which cannot be answered at the present time, concerns the extent to which the individual fragments may lose angular momentum through torque transmitted from that part of the interstellar magnetic field trapped inside the fragments when they are formed. No studies have yet been made of stellar models in this earliest stage of contraction; we do not know how fast they contract, nor whether they contain large internal convection zones, which would tend to scramble the internal magnetic field and isolate it from that of the surrounding medium. Perhaps it is best to proceed with two alternate assumptions: that during early contraction phases the protostars (1) lose *most* of their angular momentum, or (2) lose *none* of their angular momentum.

HYDROSTATIC EQUILIBRIUM

Once a protostar is formed, an approximate condition of hydrostatic equilibrium will be set up throughout. This means that there will be a central concentration of mass in the protostellar

body, with the central density becoming considerably greater than that in the outer layers. Again applying the principle of conservation of angular momentum, we would conclude that the central region of such protostars would rotate at a faster rate than the outer regions. However, such a differential rotation throughout the protostar cannot long be maintained owing to the action of the magnetic field trapped inside the protostar. The lines of force of the magnetic field stay well glued to the interior gas of the protostar, regardless of low temperature. This condition is caused by the persistent state of ionization being maintained by radioactive substances contained in the gas (particularly the radioactivity of the isotope K^{40}). The differential rotation in the protostar draws out this magnetic field into a spiral pattern, thus crowding the magnetic lines of force closer together, and increasing the energy in the magnetic field. This additional magnetic energy can be obtained only at the expense of the differential rotational energy, and hence differential rotation ceases. We may therefore regard the protostellar bodies as being rigid rotators during the course of their contraction.

The author has attempted to determine, in a rough way, the behavior of a contracting protostar by considering how polytropic spheres contract. A polytropic sphere is a spherical mass of gas in which the pressure is related throughout to the density by some given power law. Such polytropes have been studied in considerable detail since the latter part of the 19th century, and have provided useful approximations in some cases for the conditions to be expected in a stellar internal. If, throughout the polytrope, we have a relation between the pressure P and the density ρ given by

$$P = K\rho^{(n+1)/n},$$

where K and n are constants, then n is called the index of the polytrope. Two values of this index have been used in this study to represent probable extreme conditions between which the internal density distribution of the protostars may fall. In one extreme we let $n = 1.5$, corresponding to a sphere in which there is convection throughout. In such a sphere, the central density is approximately 6 times the average density. In the other extreme, the index $n = 3$ corresponds to a stellar model in which energy transport is principally by radiation. In such a sphere, the central density is about 54 times the average density.

As the protostar contracts, half of the gravitational potential energy released will be stored as internal heat, and the remaining half will be radiated away from the surface. Indeed, the rate at which this excess energy can be radiated from the surface governs the rate at which the protostar can contract. Until more realistic models of the protostar are constructed, we cannot say how rapid this contraction would be.

As the contraction proceeds, the gravitational potential energy being stored as internal heat results in an increase in the central temperature. This temperature is probably a few hundred degrees Kelvin when the protostar is first formed as a fragment, and increases in proportion to the decreasing radius of the protostar.

When the central temperature has risen to the vicinity of 1800°K, some rather interesting phenomena take place. By far the most abundant constituent of the gas is hydrogen, which will be in the molecular form at such low temperatures. When the temperature rises toward 1800°K, however, not all of the remaining half of the gravitational potential energy is radiated away from the surface of the protostar; some of it goes into the excitation of higher rotational and vibrational bands of the hydrogen molecules. Furthermore, at 1800°K, the dissociation of molecular into atomic hydrogen begins; and this too occurs at the expense of the remaining half of the gravitational potential energy. Indeed, the demand for energy to dissociate the hydrogen soon becomes so great that the temperature at the center of the protostar stops rising as rapidly as one might expect. Far more energy would be required to dissociate all of the hydrogen throughout the protostar than is possessed by its gravitational potential energy when the central temperature has reached 1800°K. Hence, the central temperature cannot increase appreciably beyond 1800°K until a very large shrinkage of the protostar has occurred and the excess gravitational potential energy has gone into the dissociation of the hydrogen molecules. Under these circumstances, the protostar can no longer remain approximately in hydrostatic equilibrium,

but must undergo a *collapse* in which the gravitational potential energy necessary to dissociate molecular hydrogen is rapidly released.

For a sphere of polytropic index $n=1.5$, the relation between the mass, radius, and temperature is

$$M/R=1.73\times10^{-5}\ T/\mu,$$

where M is the mass in solar units, R is the radius in astronomical units, T is the temperature in degrees Kelvin, and μ is the mean molecular weight of the material. For solar material containing hydrogen molecules, $\mu=2.5$. With $T=1800°$K and $M=1$, we have $R=80$ astronomical units. For a polytropic sphere of index $n=3$, the similar relation is

$$M/R=1.093\times10^{-5}\ T/\mu.$$

With $\mu=2.5$, $T=1800°$K, and $M=1$, we have $R=127$ astronomical units. Thus, a protostar of 1 solar mass unit, in which hydrogen molecules are dissociated, would become unstable against collapse when its radius shrinks to about 100 astronomical units.

Let us consider a polytrope of index $n=1.5$ having no rotation. We wish to see how far the collapse of such a polytropic sphere would proceed. Let us assume its configuration to be at the stage when most of the hydrogen molecules have been dissociated throughout. It requires 4.48 electron volts to dissociate each hydrogen molecule, and hence the total dissociation energy of the hydrogen molecules throughout the protostar is 2.69×10^{45} M ergs, where M is the mass in solar units.

We wish to find out if the collapse will stop when this total dissociation energy has become equal to 7.62×10^{45} M^2/R ergs, half the magnitude of the potential energy. If the collapse ceased at that point, then, following the collapse, $M=0.353R$. We would then have $\mu=1.4$, and the central temperature would become $T=8.1\times10^4$ M/R, or 2.86×10^4 °K. However, at this central temperature, the hydrogen atoms are almost fully ionized. Considerably more energy is required to ionize a hydrogen atom than to dissociate a hydrogen molecule; thus, it is evident that the collapse cannot cease after the dissociation of the molecular hydrogen, but must continue through the ionization of the resulting atomic hydrogen.

If we were to make a similar calculation to find the polytrope configuration at the end of the hydrogen ionization stage, we would find that helium (the second most abundant element in the protostar) had become singly ionized throughout much of the protostar's central region. Helium ionization requires the absorption of yet more gravitational potential energy, and therefore the collapse will continue. In fact, the collapse must continue until not only the hydrogen, but also the helium, has become fully ionized, with each helium atom losing two electrons. This condition of instability was first pointed out by L. Biermann and T. G. Cowling,[1] but it has since received very little discussion.

We may now ask what final configuration the polytrope assumes, once all the ionization has taken place. The sum of the hydrogen molecule dissociation energy and all the ionization energies is 3.3×10^{46} M. Following the collapse, we will have $M=4.37$ R; and, since $\mu=0.665$, the central temperature will be $T=3.84\times10^4 M/R$, or 1.68×10^5 °K. Thus, one solar mass will have collapsed to a final radius of approximately 0.228 astronomical unit. In the case of the sun, this is well inside the present orbit of Mercury.

A similar set of conclusions follows for a polytrope of index $n=3$; and we would find that the final collapsed configuration was reached with $M=2.50$ R and a central temperature T of 1.52×10^5 °K. The time required for a collapse of this sort would be only a few hundred years, starting from the initial configuration.

The protostar has now reached the configuration at which F. Hoyle,[2] in a recent theory concerning the origin of the solar system, assumed that the sun became rotationally unstable at the equator owing to the conservation of angular momentum. The subsequent shrinkage of the protosun must have been accompanied by the loss of mass in the equatorial plane, in order to conserve angular momentum. Hoyle assumes that this gas was accelerated outward to larger radii by means of a magnetic interaction between the gas in the nebular disk and the sun. This interaction both slowed down the rotation of the sun and drew the released gas out into that region now occupied by the planets. The planets supposedly condensed out of this gas.

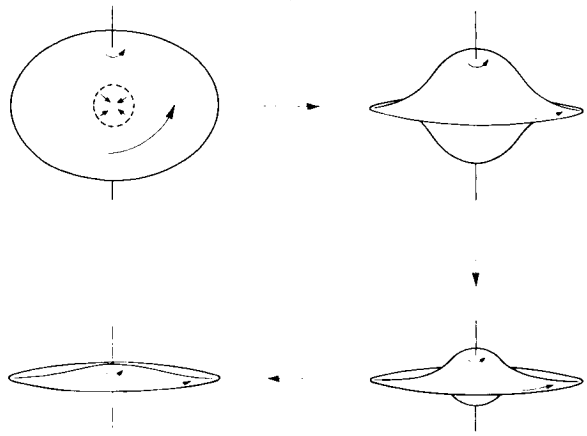


FIGURE 1.—Schematic representation of the formation of a nebular disk from a collapsing polytrope.

Let us now consider the alternative possibility in which the contracting protostars lose *none* of their angular momentum to the surrounding interstellar medium. In this case, they become unstable against the loss of mass at the equator when their radii have shrunk to a few hundred astronomical units. At the time that the central instability sets in (resulting in their rapid collapse), we would expect such protostars to be already losing mass in order to conserve angular momentum. Such protostars should remain rigid rotators until their central temperatures reach 1800°K. Once the collapse gets underway, it progresses too rapidly for the magnetic field to maintain rigid rotation throughout the collapsing polytrope: consequently, there will be a local conservation of angular momentum. Thus a given element of mass will be shed when its angular velocity becomes sufficient to permit a Kepler rotation of the element in the gravitational field of the remaining mass at smaller distances from the center of gravity.

A schematic representation of the process is shown in Figure 1, where we see a protostar with considerable distortion in the equatorial plane, owing to its rotation. When the collapse starts at the center of the protostar, mass is continually shed in the equatorial plane, and a nebular disk is formed. Some gas may remain to form a central star in hydrostatic equilibrium, about which the nebular disk will rotate.

PARAMETERS OF THE NEBULAR DISK

The author has made some numerical calculations concerning the formation of such a nebular disk. For this purpose, the mass distribution was computed in cylindrical shells about an axis of rotation through the center of the polytrope. It was assumed that the mass *of* such a cylindrical shell would be shed when the polytrope's radius became small enough—that is, when its angular velocity reached the Kepler angular velocity of a particle orbiting in the gravitational field due to the polytrope's fractional mass *enclosed by* that cylinder. The latter mass was assumed to be concentrated at the system's center of gravity. The resulting gravitational potential in which the nebular gas is presumed to move is an approximation, but it should be accurate enough for this preliminary investigation.

The first significant result to emerge from this analysis was the fact that there is no central body formed in hydrostatic equilibrium. For both classes of polytropes, the collapse leads solely to the formation of a nebular disk, in which all of the helium has been ionized. Although the density of the nebula near the center of gravity is considerably greater than that near the edge, there is no residual body at the center.

Since the collapse of the polytropes is initiated by the dissociation of hydrogen molecules, and continues owing to the ionization of hydrogen and helium (involving temperatures in a range well above 10^3 °K), it is evident that the nebular disk will be initially formed at a high temperature. The height of the nebular disk above the central plane depends upon its temperature, but the pressure at the center of the plane is independent of the temperature to a first approximation. Owing to the high temperature in the disk at its formation, the metals throughout the disk, as well as many of the other lighter constituents, will be ionized. The magnetic field, trapped in the body since its formation from the insterstellar medium, will therefore remain trapped in the nebular disk. The nebular disk will be subject to differential rotation in its own gravitational field, the inner portion rotating more rapidly than the outer. This motion will tend to draw out the magnetic lines of force, thus crowding them closer together and increasing

the magnetic field strength at the expense of the energy in the differential rotation. This is a process that tends to disrupt the nebular disk and, during the process, the nebular disk must conserve its own angular momentum. Conservation will occur if most of the mass in the disk flows inward toward the center of gravity, while the remaining mass flows outward and carries with it much of the angular momentum of the disk. From this point of view, the sun would be formed from the mass flowing toward the center of the nebular disk, as a result of magnetic friction. The precise amount of mass flowing inward to form the protosun may thus depend upon accidental configurations of the magnetic field; the effects would differ considerably from one case to another. It may be that in some cases the action of the magnetic field would result in the formation of two or more centers of condensation, causing the formation of a binary or multiple star system.

The mass deposited inside an arbitrary orbital radius is shown in figure 2. Four different initial cases are presented: polytropes of indices 1.5 and 3, and masses of 2 and 4 solar masses M_0. From this figure it may be seen that, if the initial configuration is a polytrope of index 1.5, then 1 solar mass must be collected from within roughly the orbit of Saturn. On the other hand, if the initial configuration is a polytrope of index $n=3$, then 1 solar mass would be contained inside the earth's orbit.

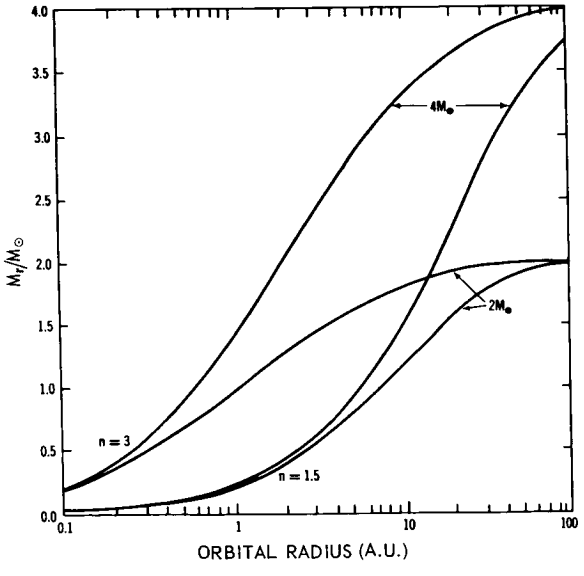


FIGURE 2.—Mass contained in nebular disks inside the indicated orbital radius.

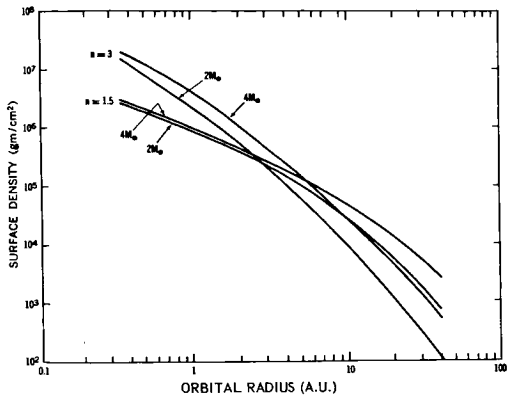


FIGURE 3.—Surface density of gas in the nebular disks.

Polytropes of considerably greater mass than that of the sun were initially assumed because, after the sun has formed from the nebula, there may be significant loss of mass. Herbig has observed very rapid rates of mass loss from stars in the later stages of their contraction toward the main sequence. Such stars are called *T* Tauri stars.

The surface densities in the nebular disk corresponding to these different assumptions are shown in figure 3. The surface density is the total amount of nebular disk mass (gm) per cm² of area in the central plane. The corresponding pressures in the disk's central plane are shown in figure 4.

Some interesting chemical and physical properties are associated with high gas temperatures and the range of pressure in the nebular gas indicated by figure 4. The gas will form at a temperature of several thousand degrees, and will cool at essentially constant pressure in the central portion of the disk. According to studies of J. A. Wood, as long as the pressure in a gas of solar composition remains above 10^{-2} atmosphere, the cooling curve will pass through a region of temperature and pressure where liquid silicates and iron will condense out. Wood has concluded that chondrules, a very common feature of meteorites, have condensed in this fashion. From figure 4, we see that a pressure of 10^{-2} atmosphere is likely

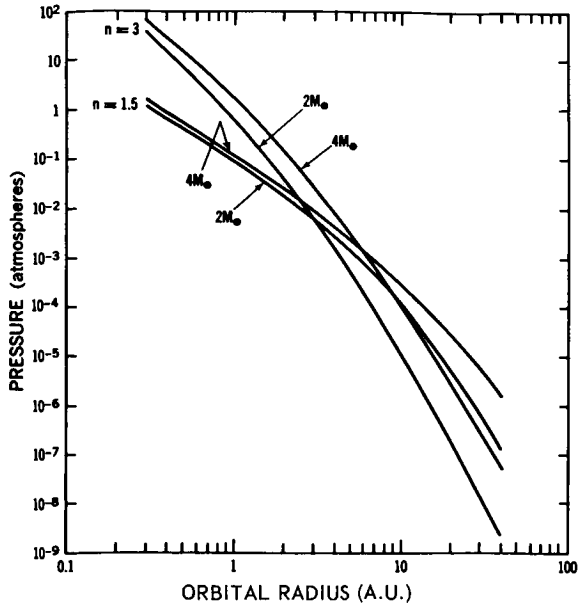


FIGURE 4.—Gas pressure in the nebular disks' central plane.

to be exceeded in the nebula out to about the orbital radii of the asteroids. This is encouraging in view of the fact that meteorites are probably asteroidal debris.

The present combined mass of the planets represents a very inefficient collection of condensable material from this nebular disk. If we consider the inner planets to be composed principally of metallic oxides, and the outer ones of metallic oxides, water, ammonia, and methane, then it appears that the present planets represent only about 1 percent of the mass that was potentially available for collection. Presumably only those bodies that grew to a sufficiently large size in a relatively short time were able to survive the dissipation of the nebula; most of the smaller bodies were swept along with the flow of gases required to form the sun. We may therefore expect that the outer portion of the solar system contains a large number of rather small solid bodies that have not collected to form planets. Such bodies may include the comets.

CONCLUDING REMARKS

It appears that the collapse of early solar evolution is a very interesting phenomenon requiring much additional study. Depending on the amount of angular momentum possessed by the earliest phase of the contracting protostar, entirely different configurations may result at the end of the collapse period; and entirely different views arise as to how the solar nebula, from which the planets condensed, was formed. Thus, studies of the problem of star formation are becoming intimately associated with the investigations of the early history of the solar system.

REFERENCES

1. BIERMANN, L., and COWLING, T. G., "Chemisch Zusammensetzung und dynamische Stabilitat der Sterne. II," *Zeit. fur Astrophysik*, **19**, 1–10, 1939.
2. HOYLE, F., "On the Origin of the Solar Nebula," *Quart. J. Roy. Astron. Soc.*, **1**(1): 28–55, September 1960.

NEUTRINO PROCESSES AND RED GIANTS

HONG-YEE CHIU*

Goddard Space Flight Center

It has been demonstrated that neutrino processes involving $(e\nu)(e\nu)$ coupling in the current theory of weak interactions are exceedingly important in later stages of stellar evolution when $T > 6 \times 10^8$ °K(Chiu, 1961).

Recent work of Adams, Ruderman, and Woo (Adams et al., 1962) has extended the importance of neutrino processes to low temperature (down to $T = 10^8$ °K) and high density ($\rho = 10^7$ gm/cm^3) when other neutrino processes become inactive. They considered the collective mode of neutrino production in a stellar plasma. Their result is summarized in figure 1. A typical value for the rate of energy loss is $dE_\nu/dt = 200$ ergs/gm-sec at $\rho = 10^7$ gm/cm^3 and $T = 10^8$ °K. This temperature-density condition corresponds roughly to that prevailing in the core of a red giant of mass $= 1.3 M\odot$ just prior to the helium-flash stage. Models for such stars have been considered in great detail by Schwarzschild and Selberg (1962).

In a previous paper the author has considered the importance of neutrino processes during the gravitational contraction phase of a degenerate stellar core. During the contraction phase, gravitational energy is released, and the temperature of the core will rise until advanced nuclear reactions start. A special case of interest occurs when the neutrino energy flux, L_ν, of the core is comparable with its gravitational energy flux, L_{core}. In less massive stars ($M <$ Chandrasekhar mass limit), the advanced nuclear reaction may not take place, and a direct evolution of stars into white dwarfs without involving supernova explosion may be possible.

Using model 16 of the red giant models ($M = 1.3 M\odot$) studied by Schwarzschild and his associates, and the results of Adams et al., we have integrated the neutrino energy flux and

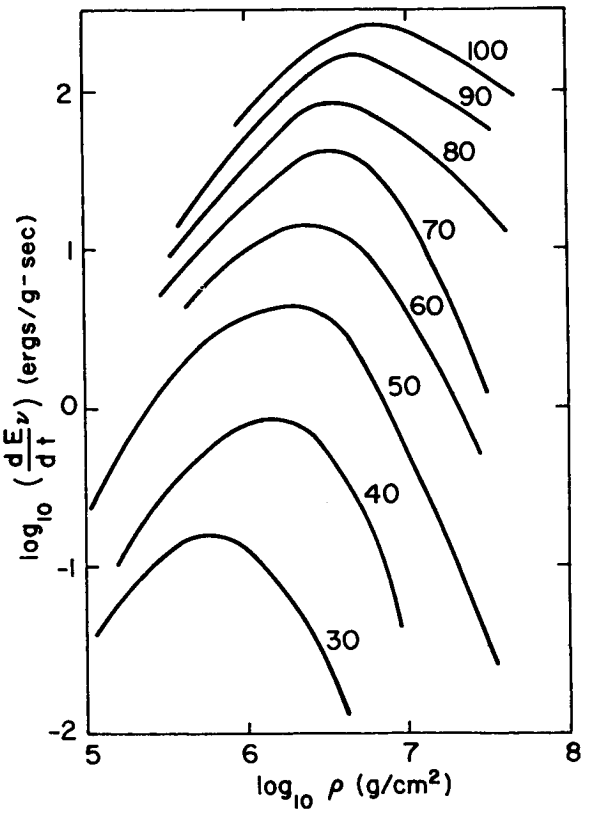


FIGURE 1.—Rate of loss of energy by neutrino emission (dE_ν/dt) as a function of the density ρ (Adams et al. 1962). The curves are labeled by the temperatures (in millions of degrees) to which they refer.

* National Academy of Science—National Research Council senior research associate with the National Aeronautics and Space Administration, Goddard Space Flight Center.

obtained:

$$L_\nu = 2.8 L\odot, \qquad L_{\text{core}} = 13.2 L\odot,$$

where the value of L_{core} is taken from Schwarzschild's model. Thus L_ν is quite comparable with L_{core}, although it is small compared with the luminosity of the hydrogen burning mantle. We therefore suggest that, in future studies of red giant models, neutrino processes such as those considered by Adams et al. (1962) may not be neglected.

It is a great pleasure to thank Prof. Edwin E. Salpeter for suggestions and discussions.

REFERENCES

ADAMS, J. B., RUDERMAN, M. A., and WOO, C. H., *Phys. Rev.*, **1962,** (in press).

CHIU, H.-Y., *Ann. Phys.*, **16,** 321, 1961.

SCHWARZSCHILD, M., and SELBERG, H., *Ap. J.*, **136,** 150, 1962.

ENHANCEMENT OF LIGHT OUTPUT FROM A SUPERNOVA

S. A. COLGATE

Lawrence Radiation Laboratory

AND

A. G. W. CAMERON

Goddard Space Flight Center

Interest has been aroused by the discovery that certain radio sources are associated with superluminous objects in distant galaxies. In particular, the source $3C$ 273 is 8–9 astronomical magnitudes brighter than a supernova explosion at the maximum of the light curve[1], and much of the emitted light appears to be synchrotron radiation[2]. Occasional flares occur in the light curve of this object, during which an excess energy of $\sim 2 \times 10^{46}$ ergs/sec is emitted within an interval of 10^5–10^6 sec[1].

Among suggestions which have recently been made to account for the large energy radiated by strong radio sources we may mention the following three:

(1) A chain reaction in a compact cluster of older stars in which one type I supernova explosion triggers another[3].

(2) Collective formation of a cluster of more massive stars with similar evolutionary timescales, which results in numerous type II supernova explosions in a limited period of time[4].

(3) Formation of a very massive star (10^6–10^8 $M\odot$), which implodes, releasing a substantial fraction of its rest-mass energy[5].

We have investigated a possible mechanism within the framework of the second of the above theories which may be able to account for the intense light flashes observed in $3C$ 273 (ref. 1). A normal type II supernova explosion would not be visible against the strong light output from the $3C$ 273 object. However, the energy in one of the light flashes is the same order of magnitude as the expected kinetic energy in the ejected envelope from a star of 10–100 $M\odot$ which undergoes a type II supernova explosion. If occasionally circumstellar gas of density $\sim 10^{-16}$ g/cm^3 accumulates in a very compact cluster of massive stars, then the interaction of the shock-ejected stellar envelope with such a gas could convert much of the kinetic energy into visible light in the required period of time. The postulated gas density is much higher than normal interstellar gas densities, which assures us that this mechanism will have only a rare occurrence, but it would not be surprising to have such a density accumulate in a region such as that under discussion.

The short time-scale of the light flashes indicates that the energy must be released in a rather small volume of space. Solar flares are believed to result from the annihilation of magnetic field energy in the solar corona; they last for $\sim 10^3$ seconds. The annihilation time varies as the square of the dimensions of the annihilation region. The amount of energy released in a $3C$ 273 light flash is so many orders of magnitude greater than that in a solar flare, and the duration is so little greater that it is extremely unlikely that there is any similarity in the processes. It seems very unlikely that the light flash energy can be stored in space

and released by any triggering mechanism; such storage would have to exceed 200 ergs/cm³ in the volume that could be illuminated by a light pulse in 10^6 sec.

Our philosophy in selecting the most likely origin of the light flashes has been to minimize the necessary energy content of the space within which the light flash is to be produced. The follow-crude analysis indicates that much of the kinetic energy of a shock-ejected supernova envelope (from a type II explosion) can be converted into visible light if a significant fraction of the envelope is slowed by collision with circumstellar gases.

Massive stars are believed to undergo implosions, that initiate type II supernova explosions, when nuclear evolution has converted their cores to iron, and the temperature has risen sufficiently to start the endothermic conversion of the iron to helium.[6] We have found that the implosion of the core continues until ordinary nuclear densities are somewhat exceeded.[7,8] In a recent theory by one of us,[9] the deposition in the envelope of the energy of the neutrino flux from inverse β-decay in the collapsing core permits a fraction ($\sim 10\%$) of the rest mass of the core to be converted into kinetic energy of expansion. The core is that fraction of the star that colapses adiabatically and in these calculations it has been $\sim 20\%$ of the total mass. Thus a total energy release of several solar rest masses ($\sim 3\times 10^{54}$ ergs) is predicted. The time of this release ($\sim 10^{-3}$–10^{-2} sec) is short enough so that a shock wave necessarily forms in the star external to the core.

Initially the shock wave will not be strong enough either to reverse the implosion of the matter or to eject it from the gravitational potential. However, as it proceeds outward it increases in strength both because of the density gradient and because of the additional neutrino energy deposition. At some point the shock becomes strong enough not only to reverse the imploding velocity but also to overcome the gravitational potential. The matter corresponding to this point asymptotically approaches final velocity after leaving the star, but the shock strength which accelerated it corresponded to the gravitational energy per gram at the time of shock passage. If nearly all the mass of the star is ejected, the mean energy of the shock becomes $\sim c^2/50$ erg/g. The corresponding mean ejection velocity for the bulk of the mass of the star would be $u_0 = 4\times 10^9$ cm/sec. Consequently, only a two-fold increase in the strength of the shock as it proceeds in the density gradient of the envelope results in matter ejected with this velocity.

The increasing strength of such a shock as it proceeds in the envelope has been previously calculated in detail.[10] The resulting velocity distribution in the nonrelativistic region can be approximated by:

$$u = u_0(10M/M_0)^{-1/5.5}, \quad M \leq M_0/10 \qquad (1)$$

where M is the integral mass external to the matter the velocity of which is u, and u_0 is the mean velocity of ejection (4×10^9 cm/sec for 100 solar masses).

If circumstellar gas of low density ρ surrounds the supernova, then half the incident shock energy $Mn^2/2$ will be converted into heat in a distance R, such that $M = 4\pi R^3\rho/3$, in a time $\tau = R/3u$ sec. The factor 3 in the denominator of the last expression makes allowance for the fact that most of the volume of the sphere is at a substantial fraction of R.

The heat deposited behind the shock may alternately be retained by the fluid or in the opposite limit may be completely radiated away. If the heat is retained, then in a spherically divergent flow the gas expands and the heat is converted into kinetic energy, leaving only a small fraction to be radiated away when the expansion causes the gas to become transparent. In the present problem we seek the opposite limit where the energy is radiated at the rate of deposition. Two conditions ensure this rapid radiation limit: (1) At thermal equilibrium, the radiation energy density Q is larger than the particle energy density, (2) the radiation mean free path (λ) is larger than the particle thermal relaxation length (l) behind the shock.

The last condition ensures that the thermal energy gradient at the shock front is at least as steep as Q/λ, and the first condition ensures that with this gradient the heat flow $Qc/4$ will be greater than the shock heating rate $Q(u_s - u_f)$, where u_s is the shock velocity and u_f is the fluid

velocity behind the shock. For u_s, $u_f \ll c$:

$$(u_s - u_f)/u_s = (\gamma-1)/(\gamma+1)$$
$$= (\text{compression})^{-1} \qquad (2)$$

so that $Qc/4$ is always greater than the shock heating rate.

We shall assume that thermodynamic equilibrium prevails, and then examine this requirement to find the conditions that optimize the visible radiant energy flux.

If we assume a uniform density distribution and a mass equal to the incident mass, the total radiation rate becomes:

$$W_t = \tfrac{1}{2}(Mu^2)/\tau = 1.5Mu^3/R$$
$$= 0.4m^{0.45}u_0^3 M_0^{0.55} \text{ erg/sec} \qquad (3)$$

which becomes as large as possible at the maximum ejected mass, $M = M_0/10$, giving:

$$W_{t\,\max} = 0.15\, M_0 u_0^3/R \text{ erg/sec} \qquad (4)$$

The maximum spectral fraction will be radiated in the visible when $3kT = h\nu = 5\times10^{-12}$ erg; or $T = 1$ eV.

The initial density ρ_0 corresponding to this radiation rate is determined by:

$$caT^4/4 = \rho_0 u^3/2$$

or:

$$\rho_0 = (1.1\times10^{-4}T^4)/u_0^3 \qquad (5)$$

The radius for equal mass at this density is determined by:

$$4\pi R^3\rho_0/3 = M_0/10 \qquad (6)$$

giving:

$$W_{t\,\max} = 0.033u_0^2 M_0^{2/3}T^{4/3} \text{ erg/sec} \qquad (7)$$

Choosing $T = 1$ eV, $u_0 = 4\times10^9$ cm/sec, and $M_0 = 100M\odot$ gives $W_t = 5\times10^{46}$ ergs/sec, mostly in the visible at a radius $R = 3\times10^{16}$ cm.

The assumption of equilibrium must be examined for these conditions.

In general, the ion slowing-down path-length due to dynamic friction with the shocked plasma electrons is small compared with the total gas thickness, provided the electron temperature is held below 5 keV by radiation. Since the relative ion energy is 8 MeV, this condition requires that the electrons radiate at least $10^3\, kT_e$. However, either the electron temperature must be in equilibrium with the radiation at a few eV, or low-energy (a few eV) bound–bound transitions must be available for non-equilibrium radiation as in the solar corona. For the latter case to be effective, the high atomic number species must not be completely stripped of bound electrons (three or more must remain bound). This latter condition for the carbon-nitrogen-oxygen group requires $T_e \leq$ 50 eV, so that again radiation in the visible requires a relatively low electron temperature. One is forced to the conclusion that should the electron temperature rise above 100 eV behind the shock front, the visible radiation would become a small fraction of the total, and indeed the remaining radiation process, bremsstrahlung with Compton scattering, is small enough so that the electron temperature would run away to such a high value, $\gg$10 keV, that the ion slowing-down length would become larger than the gas thickness, and no further shock would occur.

Therefore, we must demonstrate a self-consistent radiation loss rate large enough to maintain a low electron temperature at the heating rate of the shock [equation (5)].

The collisional shock thickness is determined by the dynamic friction between the electrons of the shocked plasma and the incident ions of the cold gas.

According to Spitzer,[11] the energy loss time of a test charge of mass M and velocity u large compared with the thermal electron velocity of an ionized plasma becomes:

$$\tau_s = \frac{u^3M^2}{(M/m)8\pi e^4 n_e z^2 z_1^2 \ln\Lambda} \text{ sec} \qquad (8)$$

For the low densities involved, $\ln\Lambda \simeq 20$, and taking $z = z_1 = 1$, the slowing-down length becomes:

$$L = \tau_s u = 5.7\times10^{-17}u^4/n_e \text{ cm} \qquad (9)$$

or:

$$L = 10^{-40}u^4/\rho_0 \text{ cm} \qquad (10)$$

For the conditions of (6) and (7) this gives:

$$L/R = 4.5\times10^{-3} \qquad (11)$$

which confirms that there is a small fractional

shock thickness, provided the radiation cooling is adequate.

The principal radiation loss from such a low-density plasma will occur due to the excitation of bound–bound transitions of the small fraction of high atomic number elements $z \geq 3$. If we assume an element composition similar to that of solar matter, then the CNO fraction, $f = 10^{-3}$, becomes the dominant radiation source. For transparent plasmas in the temperature range 5–10 eV, extensive studies of radiation loss from low-lying resonant transitions[12,13] show that the effects can be approximated by an effective allowed transition of 10 eV with a cross-section of 10^{-15} cm^2 for each atom. If the calculated radiation rate is then greater than the shock heating rate,[7] the electron temperature will fall to a low enough value so that the radiation rate is limited to partial blackbody radiation with a larger fraction in the visible.

The energy radiation rate becomes:

$$W = (\text{volume})\ n_e n_i \langle \sigma v_e \rangle (h\nu)\ \text{erg/sec} \quad (12)$$

With the assumption that nearly all the internal energy is radiated behind the shock, the shock compression would approach infinity. However, an upper limit is determined by the shock thickness, since a thinner layer would be subject to Taylor instability. The compression then becomes R/L and the 'transparent' radiation rate becomes:

$$W_R = 4\pi R^2 L f [6 \times 10^{23}\ (R/L)\rho_0]^2 \langle \sigma v \rangle (h\nu)\ \text{erg/sec} \quad (13)$$

Using condition (6), which maximizes the total available energy, and evaluating $\langle \sigma v \rangle$ for $5 \leq T_e \leq 50$ eV as 3×10^{-7} cm^3/sec results in:

$$W_R = 2.7 \times 10^{65}\, f M_0^3 / u^4 R^5\ \text{erg/sec} \quad (14)$$

Comparing this with the energy generation rate of the shock [equation (4)] shows that for radii less than the maximum R_{max} the resonant radiation can be greater than the shock rate. Therefore, the stability condition for a radiation-cooled shock can be met and the maximum radius becomes:

$$R_{max} = 3.7 \times 10^{16}\, f^{1/4} M_0^{1/2} u^{-7/4}\ \text{cm}$$

which for $f = 10^{-3}$, $M_0 = 100 M\odot$, and $u = 4 \times 10^9$, gives $R_{max} = 3.7 \times 10^{16}$ cm and a radiation rate of 5×10^{46} ergs/sec.

At least 90 percent of this radiation at the maximum radius will be in the ultraviolet, but at smaller radii the electron temperature will be reduced and a larger fraction will be in the visible. Consequently, it appears feasible to radiate an amount of energy from a shell interacting with circumstellar gas comparable with the amount measured by Smith and Hoffleit.[1]

In addition, it can be noted that the thickness of the required circumstellar gas is small enough (≤ 10 g/cm^2) so that the cosmic ray spectrum that may be produced by the shocked outer layers of the supernova[10] will be only slightly modified.

For general consistency, it is desirable to show that the model of a supernova explosion used here leads to the usual luminosity of a type II supernova at the maximum of the light curve, if there is no circumstellar gas present. The absolute photographic luminosity of a type II supernova at maximum is approximately -18 (ref. 14) corresponding to an energy emission in the photographic wavelength range of slightly less than 10^{43} ergs/sec. We may estimate the light emission from the expanding envelope in a very crude way by considering the decrease in temperature of the shock-heated envelope as it expands adiabatically, and by assuming that this temperature is also characteristic of the expanding envelope at optical depths near unity.

We have seen that the outermost layers of the pre-supernova are accelerated to near-relativistic velocities. Hence, after shock ejection, such layers will rapidly become optically thin while the temperature is very high. The bulk of the material is ejected at smaller velocities, and, as a result of the contraction of the inner parts of the pre-supernova, it will lie inside a radius $\sim 3 \times 10^9$ cm. This material will be heated by the shock wave to temperatures $\sim 4 \times 10^9$ °K (ref. 8). The energy density of radiation will then considerably exceed the thermal kinetic energy of the matter, and so the adiabatic expansion will be approximately characterized by a ratio of specific heats $\gamma = \frac{4}{3}$.

Early in the expansion, $kT/3 \gg \langle h\nu \rangle_{p\nu}$, so that the lumminosity $L \propto Tr^2$, where r is the radius of the envelope near optical depth unity. We have

$T \propto$ (density)$^{1/3} \propto r^{-1}$. Hence $L \propto r$. Later in the expansion, when $kT/3 \leq \langle h\nu \rangle_{pp}$, L becomes a sensitive function of T; at the least $L \propto T^4 r^2 \propto r^{-2}$. Hence it is evident that the maximum light emission occurs when $T \approx 10^4$ °K. This will occur approximately for a radius $r \approx 3 \times 10^9 (4 \times 10^9/10^4) = 1.2 \times 10^{15}$ cm. Hence $L_{max} \approx 4\pi r^2 \sigma T^4 \approx 10^{43}$ ergs/sec. There is great uncertainty in this figure, but it is in order-of-magnitude agreement with the observed luminosity.

This work was supported by the U. S. Atomic Energy Commission.

REFERENCES

1. SMITH, HARLAN J., and HOFFLEIT, D., *Nature*, **198,** 650 (1963).
2. OKE, J. B., *Nature*, **197,** 1040 (1963).
3. BURBIDGE, G. R., *Nature*, **190,** 1053 (1961).
4. CAMERON, A. G. W., *Nature*, **194,** 963 (1962).
5. HOYLE, F., and FOWLER, W. A., *Mon. Not. Roy. Astro. Soc.*, **125,** 69 (1963).
6. ———, *Mon. Not. Roy. Astro. Soc.*, **106,** 343 (1946).
7. CAMERON, A. G. W., *Bull. Amer. Phys. Soc.*, **6,** 285 (1961).
8. COLGATE, S. A., GRASBERGAR, W. H., and WHITE, R. H., *J. Phys. Soc. Japan*, **17,** Suppl. *A*-III, 157 (1962).
9. ———, and WHITE, R. H., *Bull. Amer. Phys. Soc.*, **8,** 306 (1963).
10. ———, and JOHNSON, M. H., *Phys. Rev. Letters*, **5,** 235 (1960).
11. SPITZER, L., Jr., *Physics of Fully Ionized Gases*, second ed. (Interscience Publishers, New York, 1962).
12. POST, R. F., Lecture presented at the Danish AEC Intern. Summer Course in Plasma Physics, Riso, Roskilde, Denmark, Aug. 1–13, 1960 (*UCRL*-6077), Riso Rep. No. 18.
13. HOFFMANN, F. W., BISHOP, A. S., and HINNOV, E., *Fourth Ann. Meeting Div. Plasma Phys. Amer. Phys. Soc.*, Nov. 28–Dec. 1, 1962, Atlantic City, N.J. Paper *F*7.
14. BAADE, W. (private communication).

THE CHARACTERISTICS OF DISTANT COMETS

BERTRAM DONN

Goddard Space Flight Center

An analysis of spectroscopic observations and colors of comets with heliocentric distances greater than 3.5 A.U. indicates that only sunlight scattered from solid grains can explain the data. Orientation and curvature found by Osterbrock in two comets are consistent with the Bessel–Bredichin theory of comet tails. Because of these characteristics, Brandt's use of these comet tails to study the interplanetary medium is not valid.

INTRODUCTION

Valuable observations of the tails of two distant comets were obtained and analysed by Osterbrock (1958). He found that the tails were intermediate between the radial direction from the sun and the tangent to the comet orbit. In his analysis, Osterbrock concluded that this result must come about through the combined action of the repulsive force of radiation pressure and the resistive force of interplanetary matter. Interplanetary material was taken to be at rest relative to the sun and to have a density of 20 protons/cm^3. Under these conditions he calculated the behavious of possible components of the tails and concluded that only the diatomic hydrides, CH, OH, NH or grains, small compared to the wavelength of visible light, would satisfy the tail observations. Radial tails were predicted for C_2, CN or dust comparable in size with the wavelength.

Although comet observations are not nearly as complete as one needs to derive a definitive model of comet behavior, they are very extensive. Many significant and suggestive observations are scattered through the literature pertaining to comets such as in the reviews by Bobrovnikov (1942), Richter (1954), Swings (1957) and current announcements. A comprehensive study and digest of all this information would be of great value for a theoretical study of comet phenomena. A true understanding of comets should be based on and explain the myriad variations of behavior among comets. Such an approach is not yet possible, nor at this stage, desirable.

A more restricted attempt based on certain generalizations of comet phenomena is possible. If this is done, one finds that the conclusions reached by Osterbrock are not consistent with the major kinematic and spectroscopic features of comets. The most favored result is that the luminosity of distant comets arises essentially exclusively from a dust scattered solar continuum and no resistive force seems necessary.

Osterbrock's conclusions concerning the interaction of the components of distant tails were subsequently used by Brandt (1961) to examine theories of interplanetary matter (Chamberlain, 1960; Parker, 1961). Brandt proposed that Chamberlain's theory was in agreement with observations whereas a large discrepancy occurred with Parker's solar wind concept was applied to comet tails. As it will be shown in this note that the comet tail model adopted by Brandt is not valid, the analysis he attempted cannot be applied to the interplanetary medium. Consequently, no choice between theories of that medium can be made at present based on observations of distant comets.

The use of comets as probes for the solar system has been pointed out on several occasions. Unfortunately, as has also been pointed out, our knowledge of comets is too incomplete to realize this possibility.

COMPOSITION

Spectroscopic observations of comets beyond 3 A.U. show a continuum only. In most instances this is definitely a scattered solar continuum

(Swings and Haser, 1957; Walker, 1958). The first spectra of Halley's Comet in 1909 (Bobrovnikov, 1931) were predominantly continuous, with possibly a trace of emission. Walker's observation of Comet Baade, 1954*h*, at 4 A.U. fell in the interval covered by Osterbrock's observations and showed no trace of emission. It is doubly significant in being the most distant spectrum of a normal comet and also in contradicting the proposal that the visible components were diatomic hydrides.

The remaining distant spectra were of comet Schwassmann–Wachmann II during active periods and all showed a reflected solar continuum only.

The following analysis shows that the coma and tail luminosity of Comet Baade cannot be explained by molecular emissions too weak to have been detected spectrographically by Walker. In particular, the diatomic hydrides cannot have contributed to the yellow or red photographs or to the visual photoelectric measurements. A dust scattered solar continuum would fit these observations and is completely consistant with all spectroscopic observations of distant comets and particularly that of Comet Baade by Walker.

Osterbrock remarks that blue, yellow or red images of Comet Baade, taken within a few days of each other, did not show any marked difference. Also Walker's (1958) three color photoelectric measurements gave a $B–V$ color index of $+0.80$, slightly redder than for the sun, $B–V=0.64$. The intensity of light in the spectrum of the comet conforms closely to what would be expected for sunlight scattered by particles of a few-tenths microns. For the three simple hydrides suggested by Osterbrock, all emissions are below 4300 Å. These would not show with yellow or red filters and they cannot have been the source of light in the coma or tail photographs.

Cometary spectra have been reviewed by Swings (1957) and by Swings and Haser (1957). The species CH, NH, OH can be identified only a little beyond 1.5 A.U. Their presence at 4 A.U. would not be expected. All ionized molecules also fade out between 1.5 and 2 A.U.

The principle emission features in comet belong to CN and C_2. The former are all less than 4300 Å except for the red CN system which would not show on Osterbrock's plates. CN has been detected at the greatest distance from the sun but disappears at about 3 A.U. The Swan bands of C_2 in comets occur only in the region from 4700 to 5450 Å with appreciable intensity. They would contribute to the yellow image but would be negligeable for the red. The C_2 emission tends to fade out somewhat before CN. Two other neutral constituents have been identified. The "4050" bands of C_3 are an important feature. These appear in the blue only and can be seen to nearly 3 A.U. NH_2 is associated with a series of strong bands between 6000 Å and 6600 Å. Weaker bands extending to 4200 Å are possibly also produced by NH_2. These emissions disappear at about the same distance as C_3.

A combination of CN, C_2, C_3, and NH_2 could account for blue, yellow, and red images. However, these molecules do not behave alike with respect to extension into the head (Swings, 1957; Swings and Haser, 1957) and the photographic images would be different. Neither does it seem too likely that they could be combined to give the $U–B$ and $B–V$ color indices found by Walker.

Not only do the ionic species, mainly CO^+ and N_2^+ disappear at distances between 1 and 1.5 A.U. but the tails they produce are long and narrow. The bands of these ions are also dominantly in the blue region (Swings and Page, 1950; Miller, 1958).

It seems certain that the direct photographs of Comet Baade cannot be produced by the known cometary emissions.

Osterbrock suggested that the tail could be composed of dust with a scattering cross section reduced by a factor of 10^4 from that for micron diameter particles. If this were the case, the scattered light would be much more intense at short wavelengths than the spectra of colour measurements show. Swings (1962) points out that dust particles in the head and in the tail may be of different dimensions. However, the light scattered by tail particles is not blue as would be the case for very small grains (Liller, 1960).

The spectroscopic evidence supports the conclusion that the observed component of distant tails is dust with dimensions of the order of the wavelength.

In his latest paper on the use of comet tails as

probes of interplanetary space, Brandt (1962) compares the deviation of tails from the radial direction for heliocentric distances less than and greater than 2 A.U. Straight tails within 2 A.U. are composed of molecular ions, primarily CO^+ whereas we have concluded in this section that beyond about 2 A.U. the dominant or sole constituent is dust. Therefore, Brandt's model for the velocity and density variation of interplanetary gas depends upon an invalid argument. It would be more appropriate to consider why the apparent composition of comets changes so markedly, as this may be related to the environment of the comet. A discussion from a different point of view is given in section IV.

STRUCTURAL FEATURES: KINEMATICS AND DYNAMICS

The other major conclusion reached by Osterbrock and adopted by Brandt was the necessity for a resistive force acting on the material of the tail. It does not appear any more necessary to introduce a resisting medium in order to explain the shapes of tails of distant comets than for comets near the sun.

The study of comet forms (Bobrovnikov, 1951) shows that tails having a continuous spectrum deviate considerably from the prolonged radius vector in the direction opposite to motion of the nucleus. For these tails, relative repulsive forces near unity are found. Therefore, near the sun tails with continuous spectra behave exactly as at greater distances.

For Comets Baade and Mrkos, tails 3′ to 8′ in length are of the order of 10^6 km. Using an average expansion velocity of 0.5 km/sec (Bobrovnikov, 1951) we obtain an interval of about 20 days to form the tail. For the angle φ between the radial direction and the line to the end point of the central line of the tail, the expression from Bessel's theory given by Wurm (1959) becomes:

$$\tan \varphi = \frac{2(2p)^{1/2}\, \xi^{1/2}}{3r\,(1-\mu)^{1/2}}. \qquad (1)$$

Here, p is the parameter of the orbit and equals $2q$, μ is the ratio of the repulsive force to solar attraction and ξ is the radial extent of the tail. According to Osterbrock's analysis, the radial and tangential components of the tails are about equal. The ξ component is then about 0.01 A.U. and inserting values for all quantities but $(1-\mu)^{1/2}$, equation 1 becomes:

$$\tan \varphi = 0.07/(1-\mu)^{1/2}. \qquad (2)$$

A value of μ of 0.99 fits Osterbrock's results. The calculation of the position angle of the tail of Comet Humason (1961e) at 5.2 A.U. by Guigay (1961) supports this result.

Osterbrock reported that the tails usually had slight curvature. This feature can also be explained by standard comet-tail theory because of the geometry of the situation which Osterbrock called attention to in regard to a different aspect. For the tails studied by Osterbrock the angle between the orbit plane and the plane of the sky through the comet was 75° at its minimum and became 90° when the earth crossed the line of nodes. The foreshortening factor fell between $\frac{1}{4}$ and 0 for the radial component. The perpendicular component on the other hand always lies nearly at right angles to the line of sight with a foreshortening between about $\frac{3}{4}$ and 1.

Figure 1 shows the shape of a tail inclined 45° to the radial direction from the sun according to equation 2. The dashed curve is the projected tail seen from the earth with about the minimum distortion expected. This was obtained by foreshortening the ξ component by a factor of $\frac{1}{3}$. As the average tail has less curvature than the

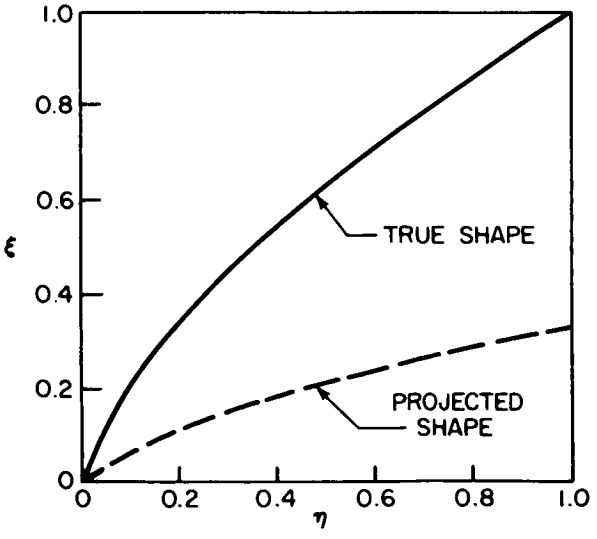


FIGURE 1.—Predicted shape of tail.

projected tail of figure 1, one expects to observe tails with only slight curvature.

In addition to projection effects, two other factors must be considered. These are: Variations of μ, and ejection of matter from the nucleus over a range of angles. When all these factors are taken into account, the problem of the shape and orientation of tails of distant comets does not appear to differ from that for nearby comets. In particular, the introduction of a resistive force is no longer necessary.

It is possible that at 4 A.U. comet tails are different. The scanty available data does not indicate such to be the case.

The variety of cometary features include besides the coma and tail, transient phenomena as halos, fans, jets and envelopes (Bobrovnikov, 1951). All these phenomena are associated with the head of a comet. With exception of the flares or halos of Comet Schwassmann-Wachmann II which occur between 5.5 and 7 A.U. these events are generally limited between 0.5 and 1.5 A.U. Although not characteristic of distant comets, the flares do tell something about the behavior of molecular species which have a bearing on Osterbrock's analysis.

At 1 A.U., CN and C_2 halos maintain their spherical shape for several days (Bobrovnikov, 1931, 1951). As Wurm (1961) points out this requires that the repulsive force on these molecules must be very small. The generally circular intensity contours for C_2 and CN in the coma (Yoss, 1953) supports this conclusion. The behavior of C_2 and CN near the sun is in contradiction with Osterbrock's calculations which indicated a strong repulsive force from radiation pressure at 4 A.U. The source of this discrepancy is not clear.

THEORETICAL CONSIDERATIONS

The preceding discussion of the limited data from spectra and colors of distant comets confirms the prevailing opinion that only scattering from solid grains is observed at large distances. Bobrovnikov (1927) reported visual observations of three bands corresponding in position and brightness with the C_2 Swan bands in Comet 1927, IV (Stearns) at 3.7 A.U. This appears to be the most remote distance at which cometary molecules have been observed. How do we explain the failure to detect gas beyond 3 or 4 A.U. as no mechanism for ejecting only solid particles from the nucleus can be devised?

We cannot conclude that no gas is present at great distances for the observations are not nearly complete or sensitive enough. But there seems to be no question that the observable ratio of dust to gas becomes very large. The very low relative intensity of the emission when the comet has receded requires a reduction in the relative density of luminous molecules for nonflaring comets for which the coma is in a quasi-steady state. The situation during a flare when again only dust is detected may be different because of the short duration of the outburst.

Oort and Schmidt (1951) have presented evidence for a greater dust concentration in their "new" comets, i.e., those having made only one or two near solar passes. A very old comet as Comet Encke has a nearly pure emission spectrum. By dividing comets into four groups according to "age" they found some indication that older comets possessing stronger emission also showed stronger brightness variation with solar distance, r. The luminosity of "new" comets varied on the average as $r^{-2.8}$ whereas for "old" periodic comets the luminosity variation was $r^{-4.2}$.

Comet emission luminosity usually is attributed to vaporization of gases and excitation by solar radiation (Bobrovnikov, 1951). Similarly, the intensity of the solar scattered continuum depends upon ejection of grains and scattering of incident light. Each of the two factors for bands or continuum varies as r^{-2} and the resultant intensity ideally goes as r^{-4}. Observations yield an average exponent somewhat less than 4, Bobrovnikov (1942) obtained a value of 3.3 for all comets without regard to spectra.

The formation of emission bands actually involves three processes. In addition to vaporization and excitation, a step involving the production of the observed radicals from stable parent species is required. The photodissociation process introduces another factor of r^{-2} causing the theoretical emission luminosity to vary as r^{-6}.

Observationally and theoretically the emission luminosity varies more rapidly than the continuum luminosity by a factor r^{-2}. This corresponds to a ratio of 16 as the comet goes from 1 to 4 A.U.

and would significantly contribute to the earlier disappearance of the emission spectrum.

A discrepancy between the observed and theoretical luminosity variations by a factor of roughly r^2 exists for both the emission and continuum spectrum. This casts considerable uncertainty on the validity of this interpretation and poses a problem for interpretations of comet luminosity in general. Without, however, more complete observational data, including photometric studies of emission bands and continuum over much greater ranges of distances than are currently available one cannot say how real the discrepancy is.

What is definite is the decrease of the emission spectrum at a more rapid rate than for the continuum. No conclusions can be reached until systematic observational studies of distant comets have shown what effects one is trying to explain.

Until further progress has been made, the use of comets as probes of the solar system will remain very restricted and conclusions derived thereby, uncertain.

REFERENCES

Bobrovnikov, N., *Ap. J.*, **66,** 439, 1927.

———, *Lick Obs. Pub.* **17,** 309, 1931.

———, *Rev. Mod. Phys.* **14,** 164, 1942.

———, *Astrophysics*, ed. by Hynek, McGraw-Hill, N.Y., 1951.

Brandt, J. C., *Ap. J.*, **133,** 1091, 1961.

———, *Icarus*, **1,** 1, 1962.

Chamberlain, J. W., *Ap. J.*, **133,** 675, 1961.

Guigay, G., I. A. U. Circulaire, 1782 Bureau Centraldes telegrammes astronomiques, Observatory Copenhagen, 1961.

Liller, W., *Ap. J.*, **132,** 867, 1960.

Miller, F., *Pub. Ast. Soc. Pac.*, **70,** 279, 1958.

Oort, J. H., and Schmidt, M., *B.A.N.*, **11,** 259, 1951.

Osterbrock, D. E., *Ap. J.*, **128,** 95, 1958.

Parker, E. N., *Ap. J.*, **132,** 175, 1960.

Richter, N., Statistik und Physik der Kometen, Johann Barth, Leipzig, 1954.

Swings, P., *Vistas in Astronomy*, **2,** 959, ed. by A. Beer, Pergamon Press, London, 1957.

———, Paper presented at Interdisciplinary Conference on Electromagnetic Scattering, Potsdam, N.Y., August 1962.

———, and Haser, L., An Atlas of Representative Cometary Spectra, Liege, 1957.

———, and Page, T., *Ap. J.*, **111,** 530, 1950.

Walker, M., *Pub. Ast. Soc. Pac.*, **70,** 191, 1958.

Wurm, K., *Handbuch d. Physik*, ed. by S. Flugge, **52,** 465, 1959.

———, *Mem. Soc. Roy. Sci. Liege*, **20,** 369, 1961.

Yoss, K., *Mem. Soc. Roy. Sci. Liege*, *4th* Ser., **13,** 72, 1953.

THE ORIGIN AND STRUCTURE OF ICY COMETARY NUCLEI

BERTRAM DONN

Goddard Space Flight Center

The composition and structure of icy conglomerate comet nuclei formed by an accumulation process in the primordial solar nebula is examined. In addition to the constituents usually proposed more complex and reactive molecules are expected. A significant proportion of compounds stable only at low temperatures should occur also. The mixture of ices and meteoric matter would form an aggregate with a density of a few tenths of a gm per cm^3. An H_2O-ice nucleus is stable to radii of 10 km. For larger objects compression begins to occur at the center causing the structure of small comets and the outer regions of large comets to differ from that of the interior. Cometary meteor showers suggest that such differences do exist.

NUCLEAR MODEL

Many general properties of comets can be accounted for by the icy conglomerate model for the nucleus described in considerable detail by Whipple (1950, 1951). As originally developed, this model cannot explain the appearance of radicals and ions as close to the nucleus as they are observed (Wurm, 1962). This is a major defect but it can conceivably be explained by a revised composition (Haser, 1955; Donn and Urey, 1956; Donn, 1960) or by some type of corpuscular or plasma interactions (Wurm, 1962; I.A.U., 1962).

In addition, the varied appearance of different comets under similar conditions suggests that considerable variation of the structure or composition of the nucleus exists. With this possibility left open, we examine the origin, internal structure, and development of an icy conglomerate nucleus of the type proposed by Whipple. The preservation of comets is assumed to be by the distant comet cloud model of Oort (1950) or some variation of it.

The nucleus has a diameter of the order of a few kilometers and is a more or less spherical block consisting of about 75% condensed volatile material of the light elements, hydrogen, carbon, nitrogen, and oxygen. The remainder is composed of nonvolatile particles. These are some mixtures of metals (iron, nickel), metallic oxides and silicates, and possibly carbon.

The detailed chemical composition and physicochemical processes of comets are a complex problem which will be treated separately from the physical aspects of the nucleus which are the concern of this paper. The assumed elemental composition represents the relative cosmic abundances of Suess and Urey (1956) modified by a large deficiency of those gases which under the conditions that prevailed prior and during the time of formation, could not form condensable compounds (Brown, 1952).

Urey (1952) has discussed the chemical equilibrium in the solar nebula for temperatures of 300° and 1200°K. His thermodynamic equilibrium will be modified by two factors: (1) Chemical effects of energetic radioactive decay products and energetic solar corpuscles (Hoyle, 1960; Fowler *et al.*, 1961); (2) the probable very low temperature of formation because of shielding of solar radiation by the condensed grains (Levin, 1957; Opik, 1962). The first factor will dissociate molecules producing various radicals and the second will result in a complex and partly nonequilibrium composition in the solid (Thrush, 1960; McGee, 1961; McGee and Martin, 1962; Papazian, 1960).

Table I shows the adopted composition. It is

TABLE I.—*Proposed Chemical Composition of Nucleus*

Material	Constituents	Concentration
		Percent
Saturated volatile	H_2O, CH_4, NH_3	40
Unsaturated (reactive)	H_2O_2, N_2H_4, NH_2OH, HCN	30
	C_2H_2, C_2H_4, more complex organic molecules inorganic and organic compounds stable at low temperatures only.	
Inert, nonvolatile	Silicates, metallic oxides, reduced metals (?) carbon grains (?)	30
Free radicals	H, NH, OH, CH, CH_2, etc	1

similar to one proposed previously (Donn and Urey, 1957) except that the concentration of trapped radicals is reduced to a very low value in line with experimentally measured concentrations (Pimentel, 1960; Brown et al., 1962; Florin et al., 1962). The concentrations of all are estimates. They are dependent upon the effectiveness of the two factors previously mentioned as well as the temperature, pressure, and other details of formation.

Whipple proposed a low density for the nucleus. In view of the fragile character and low density of cometary meteors (Jacchia, 1955; Opik, 1958) a similar structure for the icy matrix in which the meteoric particles are embedded is required. The embedding and the preservation of dust-ball meteors in a dense icy mass is an unlikely process.

Levin (1962) has suggested certain problems with this picture of the comet-meteor structure. However, his difficulties are eliminated by the mechanism for the formation and aggregation of solid particles described by Donn and Sears (1963). Their mechanism for the origin of comets is developed further in the next section.

The reactive composition of table I requires a low temperature to prevent rapid reaction of unstable species. The structure of the nucleus is also very dependent upon the temperature. It was pointed out (Donn and Urey, 1957) that because of the large proportion of volatile material, nuclear temperatures would be low. Surface temperatures are probably below 150°K, well within 1 A.U. Detailed calculations of temperatures of various kinds of ices have since been published (Squires and Beard, 1961; Watson et al, 1963). Interior temperatures tend to be more representative of aphelion distance (Minnaert, 1948) and therefore always well below 100°K even for short period comets.

An increase of temperature from any source, e.g., solar electromagnetic or corpuscular radiation, would result in decomposition of the low-temperature-stabilized molecules. The fragments would react in various ways releasing additional energy and causing further heating. This is a modification of the free radical mechanism proposed for cometary flares (Donn and Urey, 1956).

Reactions may take place in the surface layer of the nucleus or in the voids of the porous, low density matrix. When the reacting species emerge into the coma, the density becomes very low and all reactions are "frozen." A large proportion of radicals characteristic of flames (Gaydon, 1960) thus may persist in the coma.

Reactions of frozen, unstable molecules in sufficient magnitude causing the two above processes to occur could account for some of the problems of cometary phenomena. These and related matters will be examined in detail in a subsequent paper.

THE ORIGIN OF ICY COMETARY NUCLEI

Ter Haar (1948) and Urey (1952) found that condensation of all known cometary radicals could have been incorporated into the comet as stable molecules at 100 A.U. from the sun. The reduced temperature pointed out by Levin (1957), Urey (1958), and Opik (1962), would permit cometary material to condense much closer. Urey and Opik concluded that hydrogen condensed at Jupiter's distance.

In order to account for the continued existence of comets Oort (1950) proposed a cloud of comets at 10^5 A.U. These formed near the sun and were transported to the distant cloud by a combination of planetary and stellar perturbations. An essential feature of the present paper is the acceptance of the idea that comets could form during the development of the solar system. It should be noted that if comets could form near the sun, they could do the same around other stars where, however, the details of formation and consequently their structure and composition would vary.

Donn and Sears (1963) pointed out that the irregular filamentary structure expected for crystals grown in space would lead to rapid growth of low density aggregates. Drag by the gas would reduce impact velocities to meters/sec, facilitating the sticking of particles upon collision.

Low velocities of impact were also required for the composition and low density of the nucleus. If comets contain material as volatile as nitrogen and methane (melting points 63°K and 89°K, respectively), and if the temperature in space were 50°K, about 1 cal/gm would be required to melt the particle. This corresponds to velocities of 0.1 km/sec which is a lower limit as only part of the energy of a collision goes into heat. The vapor pressure of all molecules of interest is at least several mm of mercury at their melting points and they would vaporize rapidly if heated near their melting point. Temperatures near melting would have considerably changed the structure in addition to causing large loss of material.

A velocity of 0.1 km/sec corresponds to about 200 miles/hr. Experience with hurricanes and tornadoes demonstrates that such velocities would produce a highly compact nucleus without the necessity of melting. Winds of 50 miles/hr or 0.025 km/sec produce compacted snow with densities of 0.3–0.5 gm/cm^3 (Seligman, 1936). As long as no melting occurs frozen gases generally may behave similarly to snow (an experimental study of this subject is being planned). Consequently, accumulation velocities as low as deduced above indicate an accreted icy mass will have an initial density of about 0.3 gm/cm^3 or less.

A high gas abundance was certainly present during the early stages of accumulation and the condition for low impact velocity of grains or small aggregations existed. Consequently the low-temperature stable species included in table I would not be destroyed during the accumulation.

The hypothesis of this paper concerning the composition, density, and mechanism of formation of comets appears to be self-consistent.

According to the mechanism of Donn and Sears, many cometary nuclei formed. A large fraction went to form the giant planets (Opik, 1962; Whipple, 1958) and some may have contributed to the terrestrial planets also. As planetary masses formed, large perturbations of unaccreted cometary nuclei occurred. Oort's preservation process then ejected a large number from the solar neighborhood which eventually found their way into the distant cloud. Another large fraction was ejected from the solar system itself.

It should be noted that Lyttleton's (1953) mechanism of capture from an interstellar cloud could possibly develop along the lines described by Donn and Sears with the separate grains collecting into larger aggregates. This would lead to icy nuclei and remove some of the major difficulties with Lyttleton's suggestion (Whipple, 1961).

STRUCTURE OF THE NUCLEUS

Structural characteristics of condensed gases other than water vapor or of a mixture of several condensed gases are not known. However, for an icy H_2O nucleus, results of glacier studies are available and may be applied to the comet problem. If we assume that a mixture of ices generally behaves similarly to ordinary snow and ice, we can obtain some idea of the structure of cometary nuclei.

The central pressure P_c, surface gravity, g, and velocity of escape, V_e, of a sphere of radius R and mean density $\bar{\rho}$, are given by:

$$P_c = \tfrac{2}{3}\pi G\bar{\rho}^2 R^2 \qquad (1)$$

$$g = \tfrac{4}{3}\pi G\bar{\rho}R \qquad (2)$$

$$V_e = (\tfrac{8}{3}G\bar{\rho})^{1/2}R \qquad (3)$$

Table II shows the three quantities for nuclei of mean density unity and $R = 1$, 10, and 100 km.

It is of interest to note that the escape velocity, for a 100 km diameter nucleus with $\bar{\rho} = 1$ gm/cm^3,

TABLE II.—*Mechanical and Dynamical Properties of Comets*

R (km)	P_c (dynes/cm^2)	g (cm/sec^2)	V_e (km/sec)
1	1.4×10^3	0.02	8×10^{-4}
10	1.4×10^5	0.2	8×10^{-3}
100	1.4×10^7	2	8×10^{-2}

is just at the border line for causing some fusion and at the point where a low density snow does not form because of the impact velocity.

Snow densities under various conditions of packing have been determined in alpine glacier studies (Seligman, 1936, 1941) and antarctic snows (Crary, 1961).

The usual density of newly fallen snow is 0.03–0.05 gm/cm^3 (Seligman, 1936). Snow which has fallen at very low temperatures (−15°C) and in windless conditions can have a density as low as 0.01. The lowest reported density has been 0.004 (Church, 1942). Under terrestrial conditions the density increases through settling, to 0.02–0.3 gm/cm^3. Wind also packs snow to a density of about 0.3.

As the snow becomes further transformed into ice, the variation of density with depth has been determined by Seligman (1941) and his coworkers for a temperate, alpine glacier, and by Crary (1961) for an antarctic snow field. The latter's density-depth curve has been smoothed and converted into a density–pressure relation. This is shown in Figure 1. Antarctic temperatures were always below about −20°C whereas melting occurred during the summer months in the glacier and the temperature was always near the melting point at a depth of 15 m. Near the surface, densities in the two cases were similar, about 0.3 gm/cm^3. The alpine densities became slightly greater at greater depths, probably because of higher temperatures.

The principal mechanism of density increase in the early stages, especially when no melting occurs, is sublimation and recrystallization of the snow (Seligman, 1936, 1941), a process which must also take place in comets over long periods of time. Evaporation from sharp corners and points, with recondensation on less highly curved surfaces preferentially occurs. The pointed, fluffy snow crystals transform into hexagonal, transparent icy crystals with concomitant reduction in volume and increase in density. The final transition to glacier ice requires pressure and flow of the crystals. The resulting pressure-density curve depends upon the extent of trapped air and any applicability to cometary problems vanishes.

A comparison of Fig. 1 with table II shows that for an assumed mean density of 1 gm/cm^3 and R less than 10 km, the central pressure is only sufficient to produce a density of 0.45 gm/cm^3. As initial densities of 0.3 gm/cm^3 would occur by accretion, a high density icy nucleus would not form.

Studies of natural snow deposits do not suffice to determine the behavior of extremely loosely packed material because wind packing increases the density to about 0.3 gm/cm^3 by the time any appreciable depth has accumulated. The lowest densities in figure 1, therefore, exaggerate the effect of pressure. The results deduced above are upper limits for densities or lower limits for radii. For densities above about 0.4 gm/cm^3 the curve should be independent of initial density and conclusions concerning larger nuclei are not affected by the low pressure uncertainty.

The postulated actual nucleus is composed of one-fifth nonvolatile meteoric material. If this consists of compact grains of silicate minerals and metallic oxides, their mean density would be between 3 and 4 gm/cm^3, and the bulk density of the nucleus averages out to 1 gm/cm^3. For dustball meteors densities are about the same as for the volatile nuclear matrix in which they are

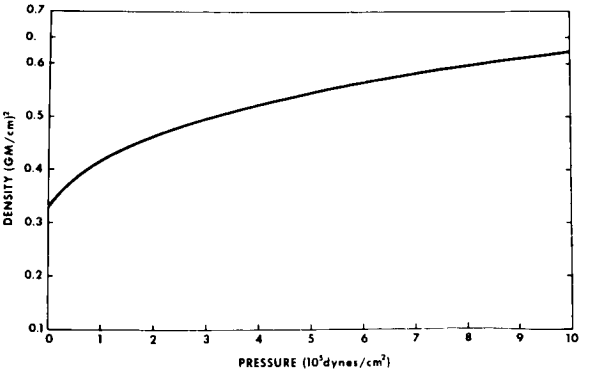


FIGURE 1.—Compression of snow.

embedded, and the overall density of the nucleus remains low.

The combination of figure 1 and the central pressures of table II suggest that compression of the porous structure of the nucleus begins to occur for diameters between 10 and 100 km. A number of physicochemical effects should occur as this happens. It appears most interesting to follow this line of investigation more fully but only a few remarks will be made here.

Quantitative differences in the behavior of nuclear structure with pressure will result when correct density-pressure relations are known for the constituents of the nucleus. However, the qualitative behavior will not change. With unstable molecules, the thermal and chemical effects suggested below may have pronounced effects.

Whipple (1960) has commented on certain observed differences among shower meteors (Jacchia, 1956). There are indications that large, old comets have less fragile or larger meteoroids associated with them, than do new comets or small comets. This is suggestive of building up of meteorids in the interior of comets and for particles from the core of large comets to be stronger as well as larger than the general meteoric particle. Whipple (1961) also proposes a solution to the anomalous increase (Oort and Schmidt, 1951), of the observed gas/dust ratio in older, periodic comets compared to "new" comets. The hypothesis is that the small grains tend to aggregate into larger ones in the interior of comets, with a concomitant reduction in light scattering cross section. This reduces the continuous solar spectrum relative to the molecular emission spectrum.

The density–pressure curve of figure 1 for condensed water can be expected to resemble this relation for other substances. The major change should be in the pressure scale. It may be significant that for ordinary snow the behavior under pressure indicates that crushing in the deep interior, and consequently heating, should begin at diameters of the order of 10 km or somewhat larger.

Because comets accumulated under low gas pressures, unlike the case of falling snow, cometary ices may be less easily compacted. The absence of an adsorbed gas layer will allow closer contact between grains and increase the adhesion (Salisbury et al., 1963). Salisbury et al. found much stronger adhesion with powdered silicates sifted at 10^{-9} mm Hg than in air. At 10^{-5} mm the degree of adhesion was similar to that in air.

The compaction of snow may indicate a weaker structure than will result for cometary ice. Another consequence is that the behavior of terrestrial snow may be representative of more volatile substances if they accumulate in a high vacuum. Such an effect if indeed applicable to comets would make our use of data from terrestrial snow deposits more valid for the cometary nucleus.

A diameter of 10 km, the maximum size of a stable nucleus deduced above, is the estimated present diameter of comets Halley and Encke. Their original nuclei must have been considerably larger (Whipple and Hawkins, 1959). Whipple (1961) suggests a factor 10 in mass for the original nucleus of Halley's comet. We can tentatively conclude that for bright, and hence, large comets, the interior of the nucleus does not resemble the outer fluffy structure of comets. The compression will have heated the core, modifying the composition of the volatile matter and the structure of the nucleus. This would be the first stage of the transformation from cometary into asteroidal objects as the mass accumulated and internal pressures increased (Whipple, 1958; Ringwood, 1959). The continued growth and transformation into giant planets has been suggested by Opik (1962), and Whipple (1958).

After a number of returns to perihelion the outer, uncompacted material will be lost and the denser, partly transformed ices exposed.

A quantitative extension of these descriptional ideas appears possible on the basis of the calculations presented here. Further developments must consider in detail chemical reactions produced by the compression, and consequence of the chemical energy released (Donn and Urey, 1956). A subsequent paper will examine chemical processes and the chemical composition for icy nuclei.

REFERENCES

BROWN, D. W., FLORIN, R. E., and WALL, L. A., Formation and decay of atoms and small free radicals at low temperatures, *J. Phys. Chem.*, **66**, 2602–2612, (1962).

BROWN, H., Rare gases and the formation of the Earth's atmosphere. *In* "The Atmospheres of the Earth and the Planets" (G. P. Kuiper, ed.), Univ. of Chicago Press, Chicago, Ill. (1952).

CHURCH, J. E., Physics of the Earth. *In* "Hydrology" (O. E. Meinzer, ed.), vol. 9, McGraw-Hill, New York (1942).

CRARY, A. D., *I. G. Y. Glaciological Report No.* **5** (1961).

DONN, B., Trapped radicals in astrophysics. *In* "Formation and Trapping of Free Radicals" (A. N. Bass and H. P. Broida, eds.), Academic Press, New York (1960).

DONN, B., and SEARS, G. W., Planets and comets, the role of crystal growth in their formation, *Science*, **140,** 1208 (1963).

DONN, B., and UREY, H. C., On the mechanism of comet outbursts and the chemical composition of comets, *Astrophys. J.*, **124,** 123 (1956).

DONN, B., and UREY, H. C., Chemical heating processes in astronomical objects, *Mem. Soc. Roy. Sci. Liege (4th ser.)*, **18,** 124 (1957).

FLORIN, R. E., BROWN, D. W., and WALL, L. A., Gamma irradiation of small molecules at 4° and 77°K, *J. Phys. Chem.*, **66,** 2672 (1962).

FOWLER, A., GREENSTEIN, J. S., and HOYLE, F., Deuteronomy: Synthesis of deuterons and light elements during early history of solar system, *Am. J. Phys.* **29,** 393 (1961).

GAYDON, A. G., "Flames," Chapman and Hall, London (1960).

HASER, L., La Conservation des radicaux libres a basse temperature et la structure des noyaux de cometes, *Compt. Rend.*, **241,** 742 (1955).

HOYLE, F., On the origin of the solar nebula, *Quart. J. Roy. Astron. Soc.*, **1,** 28 (1960).

I.A.U., Symposium on structure and development of gas tails of comets, *Trans. I.A.U., XIB*, 228–233 (1962).

JACCHIA, L., Fragmentation as the source of the faint meteor anomoly, *Astrophys. J.*, **121,** 521 (1955).

LEVIN, B. J., On the character and cause of separation of molecules during planet formation, *Mem. Soc. Roy. Sci. Liege*, **18,** 186 (1957).

———, The structure of icy comet nuclei, *Astron. Zhur.*, **39,** 763 (1962); *Engl. Transl. Soviet Astron. J.*, **6,** 593 (1963).

LYTTLETON, R. A., "The Comets," Cambridge Univ. Press, Cambridge, England (1953).

MCGEE, H. A., Jr., Trapped species and chemical synthesis at cryogenics temperatures. *In* "Proc. of Fifth Intern. Symposium on Free Radicals," Gordon and Breach, New York (1961).

———, and MARTIN, W. J., Cryochemistry, *Cryogenics*, **2,** 257 (1962).

MINNAERT, M., On the temperature of cometary nuclei, *Koninkl. Ned. Acad. Wetenschap. Proc.*, **50,** 286 (1948).

OORT, J. H., The structure of the cloud of comets surrounding the solar system, and a hypothesis concerning its origin, *Bull. Astron. Inst. Ned.*, **11,** 91 (1950).

———, and SCHMIDT, M., Differences between old and new comets, *Bull. Astron. Inst. Ned.*, **11,** 259 (1951).

OPIK, E. J., "Physics of Meteor Flight in the Atmosphere," Chap. 9, Interscience, New York (1958).

———, Jupiter, *Icarus*, **1,** 200 (1962).

PAPAZIAN, H. A., Nitrogen chained compounds as intermediates in the photolysis of solid HN_3, *J. Chem. Phys.*, **32,** 456 (1960).

PIMENTEL, G. C., Radical formation and trapping in the solid phase. *In* "Formation and Trapping of Free Radicals" (A. Bass and H. P. Broida, eds.), Academic Press, New York (1960).

RINGWOOD, A. E., On the chemical evolution and densities of the planets, *Geochim. et Cosmochim. Acta*, **15,** 257 (1959).

SALISBURY, J. W., GLASER, P. E., STEIN, B. A., and VONNEGUT, B., Adhesive behavior of silicate powders in ultra-high vacuum, *44th Annual Meeting Am. Geophys. Union, April 1963, Washington, D.C.; and Proc. of Lunar Surface Materials Conference, Cambridge, Massachusetts, June 1963,* to be published.

SELIGMAN, G., "Snow Structure and Ski Fields," MacMillan, London (1963).

———, The structure of a temperate glacier, *Geographical J.*, **97,** 295 (1941).

SUESS, H. E., and UREY, H. C., Abundances of the elements, *Rev. Mod. Phys.*, **28,** 53 (1956).

SQUIRES, R. E., and BEARD, D., Physical and orbital behavior of comets, *Astrophys. J.*, **133,** 657 (1961).

TER HAAR, D., Studies on the Origin of the Solar System, *Kgl. Danske Videnskab. Selsk. Math.-fys. Medd.*, **25,** 3 (1948).

THRUSH, B. A., Radical formation and trapping from the gas phase. *In* "Formation and Trapping of Free Radicals" (A. N. Bass and H. P. Broida, eds.), Academic Press, New York (1960).

UREY, H. C., "The Planets," Yale Univ. Press, New Haven, Connecticut (1952).

———, The early history of the solar system as indicated by the meteorites, *Proc. Chem. Soc. London*, pp. 67–68 (1958).

WATSON, K., MURREY, B., and BROWN, H., The stability of volatiles in the solar system, *Icarus*, **1,** 317 (1963).

WHIPPLE, F. L., A comet Model I. The acceleration of Comet Encke, *Astrophys. J.*, **111,** 375 (1950).

———, A comet model II. Physical relations for comets and meteors, *Astrophys. J.*, **113,** 464 (1951).

———, Notes on comets, meteors and planetary evolution, *Publs. Astron. Soc. Pacific*, **70,** 485 (1958).

———, Problems of the cometary nucleus, *Astron. J.*, **66,** 375 (1961).

———, and HAWKINS, G. S., Meteors. *In* "Handbuch der Physik" (S. Flugge, ed.), vol. 52, p. 519, Springer-Verlag, Berlin (1959).

WURM, K., The UV Solar Spectrum and Comets, *Mem. Soc. Roy. Sci. Liege (5th ser.)*, **4,** 369 (1961).

PLANETS AND COMETS: ROLE OF CRYSTAL GROWTH IN THEIR FORMATION

BERTRAM DONN
GERALD W. SEARS
Goddard Space Flight Center

The application of crystal growth theory to the formation of smoke particles from the primordial solar nebula indicates that solid particles with filamentary structures would form. Such particles would facilitate successive aggregation into planets, comets, and asteroids. The difficulties associated with the aggregation of spherical smoke particles would thus be avoided.

The earth and meteorites probably formed by accumulation of solid particles at temperatures below about 600°C.[1,2] This concept has been accepted in nearly all recent attempts[3,4] to make models of the synthesis of the solar system which would fit the requirements of current theories of stellar evolution. These studies suggest that sometime during the process, conditions favorable for the formation of small solid particles occurred. The fundamental problems are the detailed mechanisms of nucleation and growth of the primary particles and the mechanisms by which these particles aggregated to build larger objects. Several suggestions have been made concerning the latter: water and ammonia near the melting point by Urey (ref. 2, p. 122), oily substances by Hoyle,[5] and unspecified surface phenomena by Wood.[6] These investigations assumed that spherical or equidimensional particles were present.

We consider the formation, structure, and properties of solid particles in the solar nebula according to crystal nucleation[7] and growth theory.[8] The particles are expected to grow as filaments[9] and then platelets[10] or combinations of both. These extremely nonspherical crystals with few crystal defects are referred to here as whiskers. We propose that these whiskers have properties that enable them to collect efficiently at first into loosely compacted aggregates resembling the lint balls that are often under beds, and then into larger objects.

Crystal formation from a supersaturated vapor phase occurs in two steps. The formation of the smallest stable volume of a crystal (about 30 molecules) is called nucleation. Subsequently, the crystal volume increases by the mechanisms of crystal growth.

Nucleation may occur either on a particle of a foreign solid (heterogeneous) or in the vapor phase in the absence of foregin particles (homogeneous). The rate of nucleation is, in either case, a very sensitive function of supersaturation.

If the saturation vapor pressure of the bulk nucleating phase is p_0 and the partial pressure of the parent vapor phase is p, the supersaturation ratio α is given by

$$\alpha = p/p_0 \tag{1}$$

and supersaturation, σ, is defined as

$$\sigma = \alpha - 1 \tag{2}$$

According to classical nucleation theory,[7] the nucleation rate, dN/dt, nuclei created per cubic centimeter per second, for homogeneous nucleation is given by:

$$dN/dt = B \exp(-A/T^3 \ln^2\alpha) \tag{3}$$

For a given nucleation process, A is constant and B is approximately constant. T is the absolute temperature.

The extreme sensitivity of nucleation rate on supersaturation or supercooling provides the

characteristic feature of nucleation. There is a critical supersaturation, α^*, below which nucleation does not occur at an appreciable rate and above which it occurs at a rapidly increasing rate. As nucleation occurs, the supersaturation is decreased and nucleation again ceases.

The growth of a nucleus occurs by the advance of the boundary surfaces. For a perfect crystal surface, continued growth requires the formation

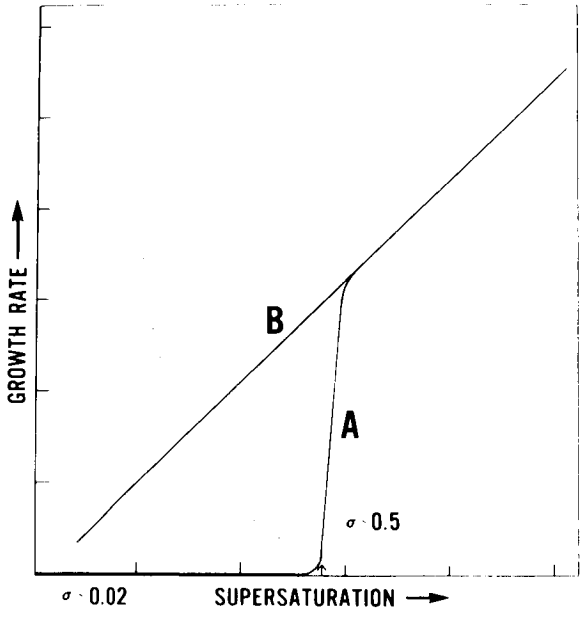


FIGURE 1.—(A) Theoretical curve for growth by two-dimensional nucleation. (B) Typical experimental curve.

of new steps on the surface[11] by a process known as two-dimensional nucleation. A sensible growth rate usually requires a supersaturation greater than 50% for two-dimensional nucleation to occur.

A schematic plot of growth rate of a perfect surface from its own vapor is shown in figure 1A. Experimental curves, figure 1B, do not display the critical supersaturation characteristic of nucleation processes.

Frank[12] pointed out that actual crystals are not perfect but almost certainly contain screw dislocations. These are defects that provide growth steps of such a nature that the completion of a surface layer perpetuates rather than removes the defects as indicated by figure 2. This eliminates the need for the formation of new steps by two-

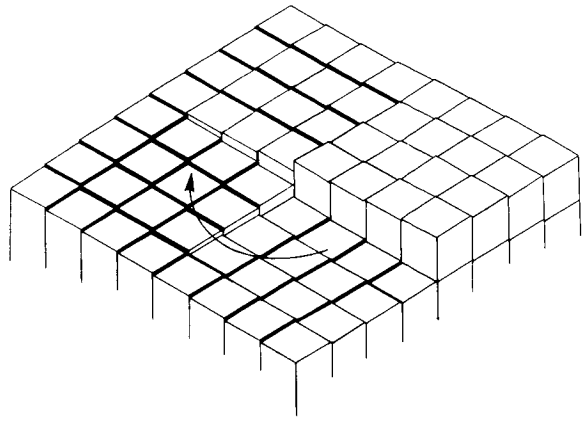


FIGURE 2.—Intersection of screw dislocation with crystal surface. The blocks represent molecules.

dimensional nucleation and allows the crystal to grow at low supersaturations. Figure 3 shows the appearance of the surface after several layers have been deposited upon it.

If a nucleus contains a single screw dislocation, it will grow at low supersaturations as a whisker[9, 13] whose axis coincides with the acting screw dislocation. If the nucleus contains a planar set of screw dislocations[14] it will grow as a very thin platelet. If it contains a three-dimensional array of screw dislocations, it will grow as an equiaxial crystal.

Since the presence of screw dislocations in growing crystallites allows equilibrium of the supersaturated vapor phase with the stable crystal phase, the dislocations guarantee that the supersaturation rapidly diminishes after nucleation. Only those nuclei containing screw disloca-

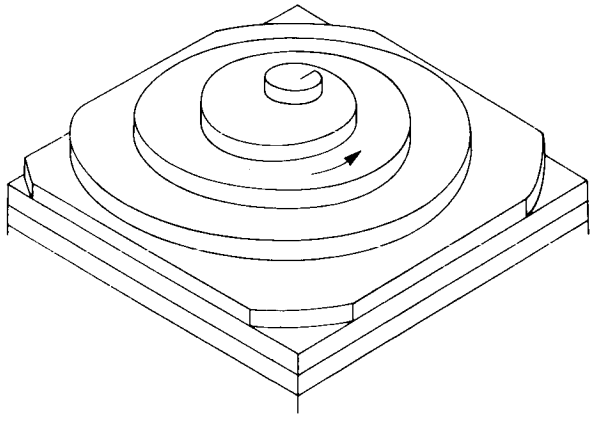


FIGURE 3.—Spiral growth at screw dislocation.

tions can grow, and the resultant smoke must contain whiskers or remain supersaturated at least to the degree required for two-dimensional nucleation. Growth at intermediate supersaturations occurs at certain re-entrant twin boundaries.

Nuclei present in the pre-planetary nebula from the initial interstellar cloud are likely to lead to screw dislocations in the growing crystal.[9,14,15] Nuclei formed at high supersaturations also contain many defects[16] and can grow as an array of whiskers,[17] Radiation damage offers a possible source of screw dislocations which could offer a path to whisker formation.

We conclude that in the solar nebula a large fraction of the small crystals will be in the form of whiskers. Since perfect nuclei will be unable to continue growing in the presence of whiskers, the condensed solids on a mass basis will consist predominantly of whiskers.

Recent experiments at Cornell University[18] have shown that powders ground to 15 μ diameter adhere readily on impact at velocities up to 15 cm/sec. They produce loosely packed structures that were called "fairy castles." These are very similar to our postulated lint balls.

The structure of filamentary particles would readily lead to coagulation by collision. The large ratio of surface to mass for whiskers compared to equiaxial particles further enhances the accumulation process. Any mechanism for forming aggregates of solid particles will be much more effective for whiskers than for spheres or equiaxed particles.

It is proposed that the whiskers collided to form lint balls. These balls provided a trap for spherical or equiaxed particles and also provided a tremendous internal area for absorption of noncondensable and slightly condensable gases.

Several properties of whiskers are of importance in the aggregation process.

1. Whiskers have strengths characteristic of perfect crystals. In general, they can be deformed 3% to 5% elastically.[19] As an example, a whisker of pure iron[20] has withstood a tensile strain of 4.9% or a stress of 1.9×10^6 psi. For comparison, an ordinary iron crystal has an elastic limit of about 5000 psi. In bending, an elastic strain of 6% has been observed[21] compared with 0.01% elastic stain for ordinary pure iron.

2. Metal whiskers would not break into fragments upon mutual collision, but would, plastically, deform and intertwine. A lint ball of whiskers should be able to catch a relatively high-speed whisker or equiaxed particle.

3. A whisker has such a high ratio of surface to volume that any heat generated during a collision would be radiated away rapidly so that evaporative losses would be held to a minimum. A metal whisker 1 μ in diameter mounted in a vacuum of 10^{-8} mm-Hg has been observed to cool at 300°C/sec from an initial temperature of 100°C at an ambient temperature of 25°C.[21]

4. The large ratio of surface to volume would provide a large cross section for whisker–whicker collisions as compared with collisions between equiaxed crystals by as much as a factor of 100.

5. The junctures of whiskers and the dislocations generated during plastic flow associated with collision would create many internal sites suitable for further growth.

Throughout the solar nebula in the region now occupied by the planets, conditions were such that nonvolatile compounds such as silicates, some metals, and metallic oxides could condense. At some distance out in the nebula, temperatures were low enough for volatile substances, for example, water, ammonia, and hydrocarbons, to condense also.[2] Shielding of solar radiation by the grains[22,23] would have brought the zone of volatile condensation in much closer to the sun. If the shielding were sufficiently effective it may have permitted hydrogen condensation as well.[22,24]

A significant fraction of whatever substances did condense would have grown as whiskers.

Because of the high temperatures that were produced in the collapsing cloud (ref. 2, p. 122) some of the condensed particles may have melted. Wood[6] has proposed that meteoritic chondrules were formed as liquid drops which then solidified or were possibly melted. Whatever processes did occur, in order to have the solid particles accrete into planets and meteorites, a majority of them must have formed as whiskers or on nuclei which grew as whiskers.

The time scale for accumulation of planets is an important consideration and requires careful analysis. Calculations have been carried out[25–27] for processes in which no mechanism of coagula-

tion was given, but high efficiency of collection was assumed. In the calculation of Chandrasekhar[27] and von Weizsacker[26] a period of the order of 10^8 years was obtained, if it is assumed that there was no significant depletion of cloud material. For a model in which the growing planets used up all the material, Schmidt obtained times of several years. Opik's studies[28] of collisions and accretion in interplanetary space are applicable and suggest times of 10^6 to 10^7 years.[29]

A cluster of grains would have a very low density, perhaps about 0.1 g/cm^3 and possibly even lower, rather than densities of 3 to 8 g/cm^3 characteristic of stones or metals, which is usually assumed. Because of the low density, particles larger by a factor of about 5 than these calculated by Hoyle[3] would be carried along by the gas. This corresponds to a minimum diameter of 500 cm.

Whatever the detailed processes, at some stage the "lint-ball" aggregates grew. Von Weizsacker, in his treatment of the problem,[26] derived turbulent velocities of 3 to 10 km/sec which were imparted to the small grains as well. The velocity must be limited to a value which would permit whiskers to coagulate. Iron whiskers could probably survive such impacts, silicates are much less likely to survive, and ice and other compounds of low boiling point would certainly not survive. A lower value seems required with a concomitant increase in time of accumulation. However, it appears that relative velocities for colliding smoke particles and small aggregates were much smaller.

The distance λ within which the smoke velocity relative to the gas is reduced in the ratio $1/e$ is given by[27,30]

$$\lambda \approx (\rho_p/\rho_g) r \qquad (4)$$

ρ_p is the particle density, ρ_g is the gas density, and r is the effective radius. This must be multiplied by the ratio of particle velocity to thermal gas velocity when the ratio is less than unity. When in addition the particle diameter is greater than the mean free path which was presumably of the order of centimeters, Stokes law must be used. These different cases will not materially affect the conclusions. If we adopt $\rho_p = 1$ g/cm^3, $\rho_g = 10^{-12}$ g/cm^3 (ref. 3), and $r = 10^{-4}$ cm, then $\lambda = 1000$ km. For lint-ball aggregates with small ρ_p, λ is still less.

The relative velocity of small aggregates was determined by the turbulent velocity of the gas within distances less than about 1000 km. A very small relative smoke velocity could not have existed or the time of accumulation would have been excessively large.

The properties of whiskers considerably reduce the restrictions on the characteristics of the solar nebula in order for accumulation to occur.

As the aggregates grew, large relative velocities could develop between large aggregates and much smaller ones. Because of the porous, low-density structure of the large objects they could absorb the smaller ones with considerable efficiency, that is with little loss of material. Collisions between roughly equal size aggregates are much more difficult to analyze but considerable fragmentation is likely. A consequence of a high relative velocity between large and small aggregates would be a tendency for the larger objects to sweep up the smaller fairly rapidly and to facilitate growth.

Small nonvolatile coagulates which formed in or moved into regions where much volatile matter condensed would have been trapped in a porous, snow-like mass. Such combinations of volatile and nonvolatile low-density aggregates would account for Whipple's icy conglomerate comet nucleus.[31] For this reason cometary grains are probably representative of the primordial condensate.[32]

When solar heating vaporizes the ices of the comet nucleus, single whiskers, small fluffy aggregates, and trapped spherulites are released. Because of their large surface to mass ratio the more whiskery particles would be more readily carried along by the cometary gases and would also undergo greater repulsion by radiation pressure than the spherical or equiaxial particles. The difference between head and tail particles, to which Swings[33] called attention, may be primarily in the shape of the grains. This could be tested by observation.[34]

Aggregates containing many grains which are released by the vaporization of the comet nucleus would spread out along the comet orbit. They would have the porous, fragile structure ascribed

to dust ball meteors associated with comets (ref. 30, ch. 9; 35; 36).

REFERENCES AND NOTES

1. BROWN, H., in *Atmospheres of the Earth and the Planets*, G. G. Kuiper, ed. (University of Chicago Press, Chicago, ed. 2, 1952); UREY, H. C., lecture, *Intern. Congr. Pure Appl. Chem., 12th, Stockholm, 1953*, p. 188; ———, *Year Book of Phys. Soc. Great Britain, 1957*, p. 14; ———, in *Physics and Chmistry of the Earth*, L. H. Ahrens, ed. (Pergamon, New York, 1957), vol. 2; LEVIN, B. Y., *Mem. Soc. Roy. Sci. Liege*, **18**, 186 (1938).
2. UREY, H. E., *The Planets* (Yale University Press, New Haven, Conn., 1952).
3. HOYLE, F., *Quart. J. Roy. Astron. Soc.*, **1**, 28 (1960).
4. CAMERON, A. G. W., *Icarus*, **1**, 13 (1962).
5. HOYLE, F., *Frontiers of Astronomy* (Heineman, London, 1955), ch. 6.
6. WOOD, J. A., *Geochim. Cosmochim. Acta*, **26**, 739 (1962).
7. TURNBULL, D., *Solid State Physics* (Academic Press, New York, 1958), vol. 3, p. 225.
8. BURTON, W. K., CABRERA, N., and FRANK, F. C., *Phil. Trans. Roy. Soc. London Ser. A*, **243**, 299 (1951); KITTEL, C., *Introduction to Solid State Physics* (Wiley New York, ed. 2, 1953).
9. SEARS, G. W., *Acta Met.*, **1**, 457 (1953).
10. NEWKIRK, J. B., and SEARS, G. W., *ibid.*, **3**, 110 (1955).
11. GIBBS, J. W., *Collected Works* (Longmans Green, London, 1928), p. 325.
12. FRANK, F. C., *Discussions Faraday Soc.*, **5**, 48, 67 (1949).
13. SEARS, G. W., *Acta Met.*, **3**, 361 (1955).
14. FULLMAN, R. L., FISHER, J. C., and SEARS, G. W., *ibid.*, **2**, 344 (1954).
15. NEWKIRK, J. B., *ibid.*, **3**, 121 (1955); KOSLOVSKI, M. I., and LOMONOSOV, M. V., *Kristallographia*, **3**, 209 (1958).
16. SEARS, G. W., *J. Phys. Chem.*, **65**, 1738 (1961).
17. COLEMAN, J. E., ALLEN, B. J., and ALLEE, B. L., *Science*, **131**, 350 (1960).
18. HAPKE, B., and VAN HORN, H., Center for Radiophysics and Space Research, Cornell University, Report No. 139, in preparation.
19. BRENNER, S. S., in *Growth and Perfection of Crystals*, R. H. Doremus, B. W. Roberts, D. Turnbull, eds. (Wiley, New York, 1958), p. 157.
20. BRENNER, S. S., *Acta Met.*, **5**, 131 (1957).
21. MURPHY, W. K., and SEARS, G. W., unpublished results.
22. OPIK, E. J., *Icarus*, **1**, 200 (1962).
23. LEVIN, B. Y., *The Origin of the Earth and the Planets* (Foreign Languages Publ. House, Moscow, ed. 2, 1958).
24. OPIK, E. J., *Mem. Soc. Roy. Sci. Liege*, in press.
25. SCHMIDT, O. J., *ibid.*, **15**, 638 (1955).
26. VON WEIZACKER, C. F., *Z. Astrophys.*, **22**, 319 (1944).
27. CHANDRASEKHAR, S., *Rev. Mod. Phys.*, **18**, 94 (1946).
28. OPIK, E. J., *Proc. Roy. Irish Acad. Sect. A*, **54**, 165 (1951).
29. ———, *Astron. J.*, **66**, 60 (1961).
30. ———, *Physics of Meteor Flight in the Atmosphere* (Interscience, New York, 1958), pp. 37–39.
31. WHIPPLE, F. L., *Astrophys. J.*, **111**, 375 (1950).
32. DONN, B., *Astron. J.*, **64**, 126 (1959).
33. SWINGS, P., in *Proc. Interdisciplinary Conf. on Electromagnetic Scattering*, M. J. Kerker, ed. (Pergamon, London, 1963).
34. DONN, B., and POWELL, R. S., *ibid.*
35. JACCHIA, L. G., *Astrophys. J.*, **121**, 521 (1955).

36. The work of G. W. Sears was under contract NAS5–1590 between General Dynamics/Electronics/Rochester and the Goddard Space Flight Center. We thank Prof. E. J. Opik for several valuable comments.

DISCOVERY OF A COMET, 5° FROM THE SUN, DURING TOTAL ECLIPSE OF JULY 20TH, 1963

FRANCOIS DOSSIN AND P. SWINGS

Goddard Space Flight Center

During the total solar eclipse, observed in the Northeastern States (Maine), a systematic search for faint comets, close to the sun, has been carried out by means of several cameras, equipped with filters. On seven plates, a nonstellar object, which exhibits cometary characteristics, has been found.

As part of the Goddard Space Flight Center, Astrochemistry Section, program of cometary studies, a systematic search for faint comets during eclipse was carried out. It has been suggested by the following considerations. It is generally observed that the luminosity of a comet is considerably increased when approaching the sun. The chance of detecting an intrinsically faint comet is the more increasing as one can observe it closer to the sun. The light of the latter, scattered by the earth's atmosphere, generally prevents astronomical observations in its direct surroundings. The use of big telescopes and large focal length allows indeed to see the stars in daylight; but for extended objects, as comets, only the ratio of luminosity must be taken into account.

Accordingly, two solutions occur for the search of faint comets near the sun: either avoiding atmospheric diffusion by means of rockets and satellites, or taking advantage of total eclipses of the sun.

Until we can use the first solution, we organized an expedition in order to take some photographs of a large region of the sky, covering more than 20° around the sun, during the total solar eclipse of July 20, 1963, observed near Caratunk, Maine (USA).

Photographic equipment was composed of: five K–37 cameras (12″ focal length; f:2.5), two K–24 cameras, modified, (36″ focal length; f:7.5), one K-40 camera (40″ focal length; f:7) and two refracting telescopes (2″ diameter) each equipped with a photographic Leica (17″ focal length; f:8 and 34″ focal length; f:16). All these instruments were equatorially mounted with a driving device to compensate for the rotation of the earth. The cameras were loaded with Eastman Kodak Super XX films. Two of the K-37's had a Kodak Wratten No. 45 filter, of which maximum transmission (35%) is on the wavelength of the cometary emission band (1–0) of C_2 (4737 Å).

Two others were fitted with a Kodak Wratten No. 22 filter. (Transmission: 80% for Na emission 5890–5896 Å).

The fifth one had no filter; other instruments were also working without filters. During the 60 seconds of totality, several pictures, with exposure times between 2 and 45 seconds, were taken, with the various cameras.

Atmospheric conditions were poor and most of the photographed field was covered with clouds. In the few clearer parts of the sky, the photographs show Venus and a few stars. In addition a slightly diffuse object, which cannot be identified with a star or another known celestial object, was found, very well marked on all the seven plates taken with one of the K–37 cameras, equipped with a blue filter K.W. 45, (C_2 region). The same part of the sky was covered with two other cameras of the same kind (K–37), one of them without filters, the second one with a yellow (Na) filter K.W. 22. On none of the four correctly exposed plates can

the object be distinguished from the background, which is quite dark.

Finally one of the f:7.5 cameras gave us five pictures, where two stars of the same region (β and κ Geminorum) appear, but the nonstellar object was not found on those plates.

To conclude: only the camera which shows together the most favorable conditions for detecting a comet with a normal emission spectrum (strong C_2, faint continuum) reveals that new object, and does it on the whole series of seven plates, taken with exposure times from 3 to 18 seconds.

Thus I propose that it is a new comet. Measures of the position were done on all plates using the only two available references: the center of the moon and β Geminorum.

The results are in good agreement and give the following approximate position: $\alpha = 7^h46^m$; $\delta = +25°23'$. The diameter of the photographic image corresponds roughly to 3′ of arc. No appreciable central condensation can be seen.

DIRECT MEASUREMENTS OF COSMIC DUST SHOWERS

M. DUBIN

NASA Headquarters

W. M. ALEXANDER AND O. E. BERG

Goddard Space Flight Center

Over the past several years direct measurements of cosmic dust by means of rockets and satellites have yielded two separate sets of data indicative of interplanetary dust showers consisting of particles whose mass is less than 10^{-8} gm. These cosmic dust showers were observed with micrometeorite detectors consisting of piezoelectric crystals and photomultipliers for detecting the light flashes from particle impacts. The Explorer I interplanetary dust stream was detected for about a 10-hour period beginning on February 2, 1958. The peak impact rate was nearly 50 times the average rate over the remainder of the 12-day period of the experiment. From the limited data it may be shown that this stream was nearly in the ecliptic in a direct heliocentric orbit. This shower does not correspond to a known annually recurring meteor shower.

A second stream of interplanetary dust has been detected with satellites and a sounding rocket. This stream appears to recur annually with a peak intensity on November 17. It was first detected in 1955 from an Aerobee rocket and was recognized in the November 1959 data from the Vanguard III satellite. Preliminary analysis of the Explorer VIII data indicates that it probably was also detected in 1960. The dust stream characteristics correspond to those of the Leonids which have a heliocentric velocity of 72 km/sec. The data supporting the detection of these two interplanetary dust streams are presented and discussed, and the physical significance of these measurements reviewed. Although direct measurements from satellites have been made over a period of about four months and during the time of other known meteor streams, only these two interplanetary dust streams have been detected thus far by direct measurements.

INTRODUCTION

The cometary origin of meteor streams is well established by the fact that the orbits of several meteor streams have been associated with comets. During the process of comet disintegration, particles are dispersed along the heliocentric orbit, and it appears reasonable to assume that for annually recurring meteor streams the age of the meteor stream is related to the uniformity in the counting rate of shower meteors. The tremendous number of meteors observed in the Draconid display in 1946 is evidence of the recent formation of the stream from the common comet Giacobini-Zinner. Dispersion of dust particles in the meteor stream from interplanetary forces resulting from the Poynting Robertson effect, corpuscular bombardment, and other radiative forces, is expected to be increasingly pronounced with decreasing dimensions of the dust particles. The extension of the range of measurements of meteoritic material to smaller masses by direct measurements with satellites may, therefore, improve our knowledge of the history and age of meteor streams.

The number distribution of meteors has been measured as a function of visual magnitude by visual and optical methods, and as a function of electron line density along the trail by radar. Browne et al.[1] have determined such distributions for the Perseids, Quadrantids and the Arietids. For the Perseids it was found, for example, that

*An exponent in the distribution function which indicates the rate of increase in number with decreasing mass.

relative to the distribution of sporadic meteors, there was a depletion of the small meteors for visual magnitudes greater than 7. Weiss[2] has measured distribution functions for the Geminids, the δ Aquarids, and the η Aquarids, as well as the sporadic meteors. Although the distribution function was found to vary for different meteor streams, only the daytime Arietids has an "s" value* equal to 2.7, compared to an s of 2.0 for sporadic meteors, and was the only shower with an s greater than 2.0.

Direct measurements of micrometeorites or interplanetary dust have been made by the United States with a number of satellites, rockets and probes. Significant data samples have been obtained on four satellites: Explorer I (1958α), launched February 1, 1958, perigee 355 km, apogee 2550 km; Vanguard III (1959η), launched September 18, 1959, perigee 509 km, apogee 3751 km; Explorer VI (1960δ), launched August 7, 1959, perigee 186 km, apogee 39,000 km; and Explorer VIII (1960ξ), launched November 3, 1960, perigee 425 km, apogee 2300 km. The results from Explorer I have been reported by Dubin[3–5] and Hibbs[6,7]. The results from Vanguard III have been reported by LaGow and Alexander,[8] and data from Explorer VIII, with distribution functions and reviews of the data have been reported by McCracken et al.[9,10]

The direct measurements of cosmic dust refer

TABLE 1.—*Impacts per Station Pass for 1958 Alpha.*

Universal time	Number of hits	Duration of pass (sec)	Station	Impact rate (hits/sec$\times 10^3$)
Feb. 2:				
1512	2	687	Santiago, Chile	2.9
1833	5	159	Woomera, Australia	31.0
1930	2	233	Quito, Ecuador	8.5
1933	1	69	Fort Stewart, Ga	14.5
2033	4	787	Woomera, Australia	5.1
2139	2	221	Fort Stewart, Ga	9.1
2238	2	691	Woomera, Australia	2.9
2338	5	192	Ehrlick, Kansas City, Mo	26.3
2339	11	366	Fort Stewart, Ga	30.0
2340	7	174	Havana, Cuba	40.0
Feb. 3:				
0043	7	832	Woomera, Australia	8.4
0132	2	213	Temple City, Calif	9.3
0138	4	259	Blossom Point, Md	15.5
0139	5	187	Ehrlick, Kansas City, Mo	26.5
0141	12	261	Fort Stewart, Ga	46.0
0142	2	7	Havana, Cuba	286.0
0258	3	180	Woomera, Australia	16.7
0334	4	213	Temple City, Calif	18.7
0338	1	126	Ehrlick, Kansas City, Mo	8.1
0344	1	71	Fort Stewart, Ga	14.0
0344	2	192	Havana, Cuba	10.4
0349	1		Antigua, B.W.I.	
0539	7	338	Jet Propulsion Laboratory, Calif	20.8
0750	2	142	Quito, Ecuador	7.0
0958	2	527	Santiago, Chile	3.8
1200	0	952	Santiago, Chile	7
1405	5	663	Santiago, Chile	7.5
Average Feb. 2–12, 11 days				1.5
Average Feb. 5–12, 8 days				0.43

to particles of masses less than 10^{-8} gm or greater than +20 on the visual magnitude scale. Daily variations in the flux of cosmic dust particles greater than an order of magnitude are often observed. There is excellent evidence that two cosmic dust showers or streams have been detected by these direct measurements. One of these dust streams, detected by Explorer I, permitted a determination of an approximate radiant; the other, however, has been detected on more than one occasion and apparently is related to an annually recurring meteor stream.

RESULTS OF MEASUREMENTS

Cosmic Dust Shower, February 1958

Evidence of a cosmic dust shower was apparent for a 10-hour period beginning on February 2, 1958 from Explorer I data. The detector was a piezoelectric crystal with an average threshold sensitivity to micrometeoroid impact of 2.5×10^{-3} dyne-sec and could detect particles of masses 8×10^{-12} gm and greater with an average impact velocity of 30 km/sec. The data from Explorer I was in real time, and it was recorded only while the satellite was over a telemetry receiving station. The data in the interval from February 2, 1512 hours to February 3, 1405 hours Greenwich time has been tabulated in table 1 for each pass over a telemetry station. Impacts were recorded for every pass, except one, during the interval of the cosmic dust shower. Included in table 1 are the station location, the number of hits during each pass, the time of each pass, and the number of hits per second.

The impact rate from table 1 has been replotted, in polar coordinates, as a function of time (figure 1). The inner cross-hatched circle represents the zero level, or zero hits per second and the next ring the average impact rate of the latter 8 days of the 12-day period that the experiment was operative; i.e., 0.43×10^{-3} hits/sec. The average impact rate over the 12-day period is the next circle and was 1.5×10^{-3} hits/sec. The plotted points are the real-time readout rates during the shower period. From this figure, it is quite evident that the impact rates during the shower period were nearly two orders of magnitude greater than the average of the latter two-thirds of the measurement period.

The dots in figure 1 represent the impact rate for all stations except for Woomera, Australia, which is represented by squares. All of the dots in figure 1 refer to receiving stations in essentially the same time zone covering about a 3-hour period. During the maximum of the shower intensity the local time for the stations in the United States and South America was early evening, which is indicative of a meteor stream in a direct heliocentric orbit. Evidence for the stream being in the ecliptic is derived from the fact that the stream was detected by northern as well as southern hemisphere receiving stations.

The Australian data from five passes provide additional evidence for a meteor stream with a direct heliocentric orbit, for it may be noted that the impact rate at Woomera as a function of time is at first high, then decreases, and again increases. The Woomera impact rate was near a minimum when the impact rate at the other stations was near a maximum. Since the longitude at Woomera differs from that of the other stations by about 180°, the earth should have shielded the satellite over Woomera from the meteor stream. The apogee (2550 km) of Explorer I was near Woomera on this date and hence, the effect of the earth's shielding at Woomera would be incomplete. Figure 1 has been shadowed to indicate approximately the extent of the meteor stream which appeared to have lasted about 10 hours. There was no known meteor stream related to the cosmic dust shower detected by Explorer I.

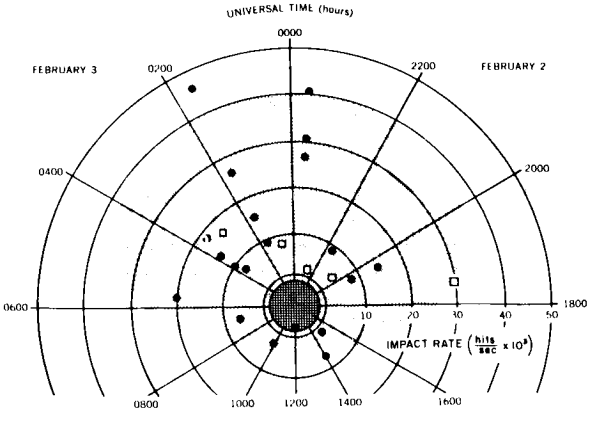


FIGURE 1.—Impact rates during the February 1958 interplanetary dust particle event.

Annually Recurring Cosmic Dust Shower in November

In November a cosmic dust shower was detected by both satellites and sounding rockets. It was first recognized from the micrometeoroid data of Vanguard III, although it was first detected in 1955 from an Aerobee rocket. The analysis of additional data from Explorer VIII is still in the preliminary stage, but it supports the fact that this shower reoccurs annually.

A number of rocket flights were carried out by Berg in 1955 and 1960, using a micrometeorite detector which detected the impact-flash of light resulting from the hypervelocity impact of the particle on an aluminum-coated lucite, or quartz surface. The results of the rocket flight in 1955 have been described by Berg and Meredith.[11]

Figure 2 schematically presents the detectors used on three separate rocket firings. The first of these firings was made with Aerobee NRL–25 on November 17, 1955, at 0215 MST. A 1P21 photomultiplier tube was used to detect the light flashes from impacts on an aluminized lucite cone with a light sensitivity of 10^{-4} lumen-sec/m^2. Assuming that 0.002 percent of the impact energy is transformed into visible light, a detector would be able to detect particles of mass greater than 10^{-13} gm. Berg found that 101 impacts had been observed on the telemetry record in the 84-second period when the rocket was above an altitude of 85 km. Below this altitude the impact rate decreased fairly symmetrically on both the upward and the downward trajectory. The area of the detector was 75 cm^2 and the impact rate was 1.6×10^2 impacts/m^2-sec.

This rather high impact rate observed on NRL–25 remained an anomaly for a number of years. This experiment was repeated by Lovering[12] in

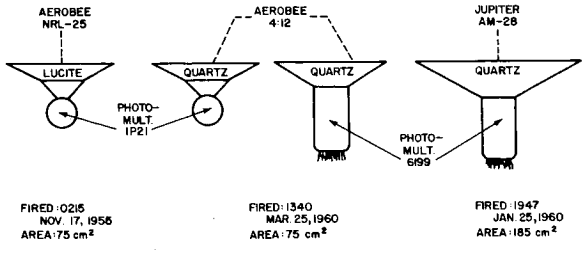


FIGURE 2.—Light-flash micrometerorite detectors used on sounding rocket flights.

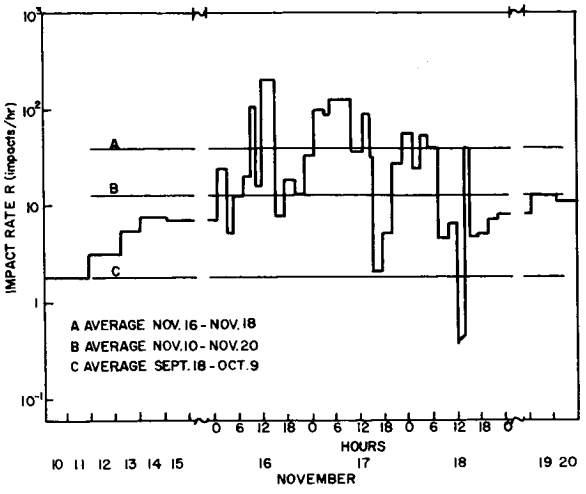


FIGURE 3.—Impact rates during the November 1959 interplanetary dust particle event.

Australia; no impacts were observed on a single rocket flight. In 1960, Berg repeated the experiment on an Aerobee NASA 4.12 and on a Jupiter AM–28. Aerobee NASA 4.12 with a 75 cm^2 detecting surface and with the 1P21 and 6199 photomultipliers was launched on March 25, 1960, at 1340 EST. One impact was observed on the 1P21 and two events were observed on the 6199. The impact rate was 1.2 impacts/m^2-sec and the exposure area-time 2.4 m^2-sec. On the NRL-25 Rocket the exposure area-time was 0.64 m^2-sec. Similarly, on the Jupiter AM–28 fired on January 25, 1960, a space-oriented detector with 185 cm^2 was exposed at 1947 EST. The 6199 photomultiplier was used with a quartz cone. Also included was a calibrated light source which gave a 50-microsecond light pulse every 26 seconds and indicated that the experiment was working correctly. The area-time of the exposure was 2.4 m^2-sec. Four events were detected which gave an impact rate of 1.6 impacts/m^2-sec. Thus, three separate rocket flights of an experiment similar to that flown in 1955 were carried out and not one of them was able to confirm the results of the 1955 flight within a factor of 100 for the impact rate.

Although it had been surmised that the rocket data could be explained by a dust shower; this hypothesis was rather clearly confirmed by Alexander[13] using the micrometeorite data from Vanguard III. The Vanguard III micrometeoroid

experiment consisted of four piezoelectric crystals with a threshold impact sensitivity of 1×10^{-2} dyne-sec. Assuming an average impact velocity of 30 km/sec, particles of mass greater than 3.3×10^{-1} gm could trigger the counter on the satellite. Data were recorded for 78 days with the total number of events approximating 5000 impacts. Real-time data were not obtained on Vanguard III except in rare instances, although the data sample was much greater than on Explorer I.

Figure 3 presents the impact rate as a function of the data from November 10 through November 20. The impact rate marked *C* is the average rate from September 18 through October 9; *B* the average from November 10 to November 20, and *A* the average from November 16 to November 18. It is quite evident that in the period from November 16–18, there are periods when the impact rate was two orders of magnitude greater than the average marked *C* (also the average of the 78-day period). It is interesting to note there were rapid fluctuations in impact rate during the period of the shower, and that several orbits showed no or very few impacts, even though extremely high-impact rates were observed for small segments of the orbit. These rapid fluctuations in the impact rate are probably real fluctuations in the spatial density of interplanetary dust particles.

The Aerobee NRL–25 was fired on November 17, 1955, at 0215. This is the same time interval of the intense impact rate as observed on Vanguard III and leads to the hypothesis that both measurements detected a shower and that this cosmic-dust shower recurs annually. From the times of the Vanguard III data and the Aerobee NRL–25, it is possible that that shower is related to the Leonid meteor stream. Preliminary analysis of the Explorer VIII data from the light-flash micrometeorite experiment using a 6199 photomultiplier with 1000A evaporated layer of aluminum indicates than an impact rate similar to that observed on NRL–25 was also observed in November 1960, during approximately the same period. However, these data are still being analyzed and are not in a form which could be used to prove or disprove the annual recurrent nature of this cosmic-dust shower. The piezoelectric experiment on Explorer VIII was oriented so as to be shielded from the Leonid radiant; and, in fact, the microphone of this spin-stabilized satellite did not detect a shower at this time.

DISCUSSION AND CONCLUSIONS

Evidence for large fluctuations in the flux of cosmic dust has been presented from the available data on direct measurements of cosmic dust using satellites and rockets. In addition to daily variations often greater than an order of magnitude in the cosmic-dust flux, two sets of conditions have occurred which may be attributed to cosmic-dust streams. Large fluctuations are also apparent in the streams themselves.

Although the total amount of data available from direct measurements is still fairly small, these two cosmic-dust streams represent a large fraction of the total number of impacts recorded on all the satellites. The total number of impacts recorded on Explorer I was 145 and the number of impacts which may be associated with the cosmic dust shower is 66—nearly half similarly, on Vanguard III, of a total number of approximately 5000 impacts, about 2800 of these impacts occurred during the cosmic-dust shower between November 16 and November 18. Even if the number of impacts recorded thus far on Explorer VIII and on all other space vehicles were included in the total, the number of events occurring in these streams represent a considerable fraction of this total. By comparison, the number of meteors in meteor streams observed by optical, visual and radar detectors is only about 10 percent of the number of sporadic meteors similarly observed. On the other hand, Gallagher and Eshelman,[14] have reported that a large fraction of the meteors observed with the highly sensitive radar equipment capable of detecting meteors of approximately visual magnitude (+14) appear to be in streams. Thus we wonder whether this supports the hypothesis that there is a continuous generation of dust by disintegration of conglomerates of interplanetary material and that the lifetime of the dust in interplanetary space is quite short—and thereby explains the relative unimportance of the space density or background density of sporadic dust particles. On the other hand, if the rate of dispersion of dust particles after disintegration from a large conglomerate were

rather rapid, and if the rate of removal of dust were slow, then the large majority of the dust particles would be isotropically dispersed; and the variations in the cosmic dust flux as detected with satellites would be small and the number of sporadic dust particles would far outweight the component of dust in showers. Additional measurements are certainly required to determine the generative and destructive characteristics of interplanetary dust.

REFERENCES

1. Browne, I. C., Bullough, K., Evans, S., and Kaiser, T. R., "Characteristics of Radio Echoes from Meteor Trails. II. The Distribution of Meteor Magnitudes and Masses," *Proc. Phys. Soc. London*, **69**(1): 83–97, February 1956.
2. Weiss, A. A., "The Distribution of Meteor Masses for Sporadic Meteors and Three Showers," *Australian J. Physics*, **14**(1): 102–119, March 1961.
3. Dubin, M., "Meteoritic Dust Measured from Explorer I," *Planetary and Space Sicence*, **2**(2/3): 121–129, April 1960.
4. ———, "IGY Micrometeorite Measurements," in *Space Research: Proc. 1st Internat. Space Sci. Symp. Nice, January 1960*, Amsterdam: North Holland Publ. Co., 1960, pp. 1042–1058.
5. ———, "Remarks on the Article by A. R. Hibbs, 'The Distribution of Micrometeorites near the Earth' ", *J. Geophys. Res.*, **66**(8): 2592–2594, August 1961.
6. Hibbs, A. R., "The Distribution of Micrometeorites near the Earth," *J. Geophys. Res.* **66**(2): 371–377, February 1961.
7. ———, "Author's Reply to the Preceding Discussion on the Article, 'The Distribution of Micrometeorites near the Earth' ", *J. Geophys. Res.*, **66**(8): 2595–2596, August 1961.
8. LaGow, H. E., and Alexander, W. M., "Recent Direct Measurements of Cosmic Dust in the Vicinity of the Earth Using Satellites," in *Space Research: Proc. 1st Internat. Space Sci. Symp., Nice, January 1960*, Amsterdam: North-Holland Publ. Co., 1960, pp. 1033–1041.
9. ———, ———, and Dubin, M., "Direct Measurement of Interplanetary Dust Particles in the Vicinity of the Earth," *Nature*, **192**(4801): 441–442, Nov. 4, 1961.
10. ———, and ———, "The Distribution of Small Interplanetary Dust Particles in the Vicinity of Earth," Paper presented at the Internat. Symp. On the Astronomy and Physics of Meteors, Cambridge, Mass., August–September 1961 (to be published in *Smithsonian Contrib. To Astrophys.*).
11. Berg, O. E., and Meredith, L. H., "Meteorite Impacts to Altitude of 103 Kilometers," *J. Geophys. Res.*, **61**(4): 751–754, December 1956.
12. Lovering, J. F., "Micrometeorite Impacts to an Altitude of 135 km," *Planetary and Space Science*, **2**(1): 75–77, October 1959.
13. Alexander, W. M., McCracken, C. W., and LaGow, H. E., "Interplanetary Dust Particles of Micron-Size Probably Associated with the Leonid Meteor Stream," *J. Geophys. Res.*, **66**(11): 3970–3973, November 1961.
14. Gallagher, P. B., and Eshelman, V. R., " 'Sporadic Shower' Properties of Very Small Meteors," *J. Geophys. Res.*, **65**(6): 1846–1847, June 1960.

AN INTERPRETATION OF BETA LYRAE

SU-SHU HUANG

Goddard Space Flight Center

To explain the β Lyrae system's peculiar spectroscopic and photometric behavior, a model in which the primary component is assumed to be the less massive of the two components has been constructed. Consequently, a theory of emission lines that predicts a redshifted peak superimposed on a broad emission line has been proposed. The observed shift in the γ-velocity of the emission peaks is thereby understood. Finally, the evolution of the component stars leading to the present situation is discussed.

A MODEL OF THE SYSTEM

In the present paper we shall present a model of the β Lyrae system in which the secondary is more massive than the primary component (Gaposchkin, 1956; Huang, 1962; Woolf, 1962). We propose that the system is similar to the Algol-type binaries. The B8 primary, like the secondary in the Algol-type variables, has completely filled its equipotential lobe, while the more massive secondary, like the primary in the Algol-type variables, is comparatively smaller in size and is surrounded by a rotating disk. The rotating disk resembles the rotating rings suggested by Joy (1942, 1947) for the Algol-type stars but is dense and therefore opaque. This disk cuts off a considerable amount of the radiation received from the secondary at all phases and obscures a part of the primary component during principal eclipse. Figure 1 illustrates two views (front and edgewise) of the model proposed here. For reasons that will be apparent later, the inclination of the orbital plane, which is assumed to be identical with the plane of the rotating disk, is not exactly 90°. Thus, when seen from the earth, the surface of the disk is projected on the celestial sphere in a significant area.

As a result of this obscuring disk, which plays an important part in the phenomenon of this system's eclipse, the relative temperature determined by the depths of both primary and secondary eclipse is no longer meaningful. This can easily be understood when we realize that the opaque disk has its own temperature.

It naturally follows from the existence of the opaque disk that the sum of radii R_1+R_2, as derived from the duration of eclipse, no longer represents the sum of radii of the primary and secondary components. Instead, it is the sum of the radii of the primary component and the obscuring disk. Accordingly, we should be able to observe, as we actually do, the B8 component during principal eclipse because the obscuring disk eclipses only the central zone of the stellar disk of the primary component. Thus the light curve will behave, as regards its duration of eclipses, like a contact binary, while actually it is not. This removes the difficulty arising from the relative sizes of the two components.

Let us now examine the size of the secondary component. Assuming the mass ratio, we can derive the masses of both components from the mass function obtained observationally. If the secondary should be a main-sequence star, we could estimate its radius from the mass. Table 1 lists the values of the radius R_2, estimated in this way for different values of $\alpha=m_1/m_2$ and for two values of the orbital plane's inclination, i. In all cases R_2/a appears to be very near to 0.1.

The radius of the primary component, R_1/a may be derived from the size of the inner contact surface, as all observational results point to the fact that it fills one lobe of the surface (Kuiper, 1941). It is also given in the table.

From the relative size of the primary and the secondary components obtained in this way, we must conclude that either (1) the secondary component is actually larger than that we have estimated or (2) the projected thickness of the opaque disk on the celestial sphere is quite large; otherwise we will not obtain a primary eclipse as deep as that observed. This conclusion may not be true if α is considerably smaller than those values listed in table 1—a situation which we will not consider seriously at present. Then, according to the *first* alternative, the secondary component is not a main-sequence star. The *second* alternative suggests that the inclination, i, differs appreciably from 90°. Both alternatives may be true. We shall return to the nature of the secondary in the last section of this paper.

The projected area of the disk on the sky depends on the extent of the disk itself, as well as on the inclination of its plane. Although we cannot know at this stage the exact size of the disk, it is apparent that it must lie within that lobe of the inner contact surface that surrounds the secondary component. Therefore, the upper limit of the disk's radius, denoted by R_2', may be obtained from the dimension of the inner contact surface and also is approximately given in table 1.

Some interesting features of the light-curve

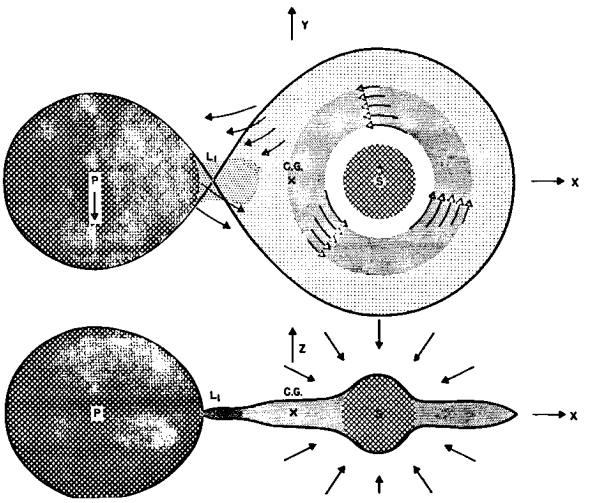


FIGURE 1.—Two views of the proposed model of the β Lyrae system. The upper one is a view of the plane of the orbit (the XY-plane), where P and S denote, respectively, the primary and secondary component. The obscuring disk is rotating differentially around the secondary within one lobe of the inner contact surface whose intersection with the plane of orbit is represented by a figure-8 curve in the diagram. The lower one is a side view of the system. The emitting atoms fall into the secondary from all directions. If the inclination of the inclination of the system is not exactly $\pi/2$, this spherically symmetric distribution of emitting atoms will produce an asymmetric profile of the emission lines, as observed, because of the presence of the obscuring disk.

TABLE 1.—*Possible Dimensions of the System*

a	R_1/a	R_2'/a*	$V_r \sin i$† (km/sec)	$i=90°$			$i=75°$		
				m_2 In $m\odot$	a In 10^{12} cm	R_2/a‡	m_2 In $m\odot$	a In 10^{12} cm	R_2/a‡
0.20	0.24	0.49	53	12.2	3.95	0.096	13.6	4.09	0.101
0.24	0.25	0.48	57	13.1	4.08	0.098	14.5	4.22	0.102
0.28	0.26	0.47	62	13.9	4.21	0.099	15.4	4.36	0.103
0.32	0.27	0.46	67	14.8	4.34	0.100	16.4	4.49	0.104
0.36	0.28	0.45	71	15.7	4.47	0.102	17.4	4.63	0.106
0.40	0.29	0.44	75	16.7	4.61	0.103	18.5	4.77	0.107
0.44	0.30	0.43	79	17.6	4.74	0.104	19.6	4.91	0.108
0.48	0.30	0.42	83	18.6	4.87	0.105	20.7	5.04	0.109

*The listed values in this column give the upper limit of the lateral extent of the opaque disk.

†The listed values in this column denote the projected rotational velocity of the primary component if rotation is synchronized to orbital motion.

‡The listed values in this column represent the ratio of the radius of the secondary to the separation under the condition that the secondary be a main-sequence star.

obtained by β Lyrae may be understood quite naturally in terms of the present model.

First, the primary eclipse is asymmetric, with the decline steeper than the rise (Stebbins, 1916). This is perhaps due to obscuration by those gases just streaming out from the primary component. They provide an opacity in addition to that produced in the gaseous disk. These gases will either join the disk or disperse away, depending on their angular momentum. Consequently, very little is left when we observe the system at phases before principal eclipse.

Second, the depth of principal eclipse varies between 0.8 and 1.1 mag in blue light, according to Guthnick (1945–46). This variation was not noticed in recent observations by Wood and Walker (1960); since their observations covered only two consecutive principal eclipses, they suggested that the variation in the depth of principal eclipse noticed by Guthnick might occur over a period of years instead of weeks. If so, the cause of the variation must be stable over a short time, say a few cycles, but unstable over a long period, say a few years or so. It follows from this reasoning that the cause cannot be a star, which should not change in a few years; neither should it be randomly flowing gases, which should change from cycle to cycle. A rotating disk consisting of gases can perhaps fulfill the stringent conditions of semistability deduced from observations. Its intrinsic thickness or its extent may vary over a period of a few years, causing a change in the depth of eclipse.

While the basic structure of the rotating disk does not change from cycle to cycle, it is reasonable to expect that there are minor fluctuations, in the density as well as in the extent, along the edge of the disk. Such fluctuations may be the cause of minor irregularities in the light-curve observed from cycle to cycle.

To test the hypothesis of the disk, we may suggest observations of the light-curve in the infrared region, where radiation emitted by the disk itself may reveal its existence.

Assuming the secondary component to be more massive, can we understand the system *without* introducing the obscuring disk? An extended scattering atmosphere that fills up the lobe of the inner contact surface around the secondary may explain the fact that the light from the primary can be seen during principal eclipse, even though the secondary is larger than the primary. But it is difficult to see how an extended scattering atmosphere can be supported by a star whose light we do not even observe. Also, there is no spectroscopic evidence of the presence of a scattering atmosphere.

INTERPRETATION OF SPECTROSCOPIC RESULTS

Spectroscopically, Struve's (1941) interpretation can easily be incorporated into our disk model. The violet-shifted lines seen immediately after the conjunction during principal eclipse show the ejection of mass from the primary. We are able to see these lines, in spite of the presence of the disk, because the inclination of the orbital plane differs from 90°. However, the red-shifted satellite lines seen immediately before the conjunction cannot be said to be coming from the secondary, which, according to our model, does not fully fill the equipotential surface. We suggest that this system of spectral lines is connected with the rotating disk around the secondary. What we observe may be either the rotating stream that is spilled over the main disk or a stream that is diverging from the disk to return to the primary.

It is interesting to note here that, within the range of mass ratios assumed in table 1, the spectroscopic data lead to rotational velocities for the disk of the order of 200–400 km/sec at points near the inner contact surface. These values are comparable with the radial velocities of the two systems of satellite lines during principal eclipse, as the violet-shifted lines indicate radial velocities of 80–360 km/sec and the red-shifted lines 120–200 km/sec (Struve, 1957). We can easily understand why the red-shifted lines should cover a smaller range of velocities than the violet-shifted lines. While the ejection of gases naturally spreads a wide range of velocities, the gaseous particles will somewhat equalize their velocities through collisions when they move around the secondary. Furthermore, according to the present model, the stream corresponding to the red-shifted lines must be weaker than the primary stream just ejected from the primary, since some of the primary stream may coalesce

with the disk or be dissipated in other ways in the course of a revolution around the secondary. This prediction also agrees with observations (Struve, 1941).

As regards the B5 spectrum observed in this system, the introduction of the rotating disk around the secondary introduces no complication into Struve's (1941, 1950, 1957, 1948) interpretation of a shell around the entire system.

Because of the closeness of the two components, we could expect axial rotation and orbital revolution to be synchronized. But, if we should follow the conventional model of a more massive primary, we would immediately see that synchronization could not be the case (Struve, 1957), as the observed rotational velocity of the primary component is only about 45 km/sec (Mitchell, 1954; Struve, 1958), while, according to a previously accepted value of $m_1/m_2=1.5$, the synchronized rotational velocity should be of the order of 180 km/sec.

In order to explain this contradiction, Kopal (1959) suggested that synchronization had been disrupted by the star's rapid rate of evolutionary expansion. According to Kopal, axial rotation and orbital revolution are indeed synchronized before the primary departs from the main sequence. As the star expands, however, rotation slows down as a result of conservation of angular momentum.

Since the adjustment time of a stellar envelope to an external field is shorter than the orbital period, tidal bulges will move in such a way that they are always pointing to the companion star. Such a tide will perhaps induce a rotation of the surface layer not greatly different from that to be expected from the state of synchronization in a relatively short time. Therefore, it is hard to reconcile the observed rotational velocity of only 45 km/sec to the synchronized velocity of 180 km/sec, even with the assumption of evolutionary expansion. Only a reversal of the relative masses of the two components can reduce this serious discrepancy to a reasonable situation, as can be seen from table 1, where the projected rotational velocity of the primary, $V_r \sin i$, under the condition of synchronization is given as a function of α.

From the luminosity of the star we have concluded that α perhaps lies between 0.26 and 0.44 (Huang, 1962). Now, if we accept 45 km/sec as the observed rotational velocity, we find from table 1 that the deviation from synchronization is within reasonable limits and may be due to several factors. The mass ratio may be even slightly less than the suggested limit of 0.26; the observed rotational velocity may be underestimated, or, in view of rapid evolution, a slight deviation from synchronization may be intrinsic.

A THEORY OF EMISSION LINES

The density of material, although high in the equatorial plane, must decrease rapidly on both sides of the disk as we move away from the plane. Thus gas away from the plane is rare but extends a large volume. Seen from the earth, it is projected either on the opaque disk, which must be at a temperature much lower than that of the B8 component, or on the dark sky. Hence we propose that the emission lines are produced mainly by the material in this volume. Those particles that stray from the main disk most likely have small components of angular momentum in the plane of the disk; as a result, they will eventually fall into the secondary component. When the atoms depart from the disk, they have low excitations corresponding to the temperature of the disk. Once high above the main disk, they are exposed to, and consequently excited by, the radiation coming not only from the primary but also from the secondary. Thus these excited atoms will eventually emit radiation that we see as emission lines. Since atoms are excited at places high *above* the disk, emission takes place dominantly when the atoms are falling down toward the secondary star.

To derive the profile of the emission lines, let us first consider an ideal case, in which the emitting atoms have a spherical symmetry with respect to the secondary star and fall to the star with a constant speed V. We choose the center of the secondary component as the pole of a spherical co-ordinate system with the polar axis in the direction of the line of sight, and we assume that the natural width of the emission line is infinitesimally narrow, so that the broadening of the line is solely due to the Doppler effect. As the radial velocity of an emitting atom at any point (r, θ, φ) is $v=V\cos\theta$, which is inde-

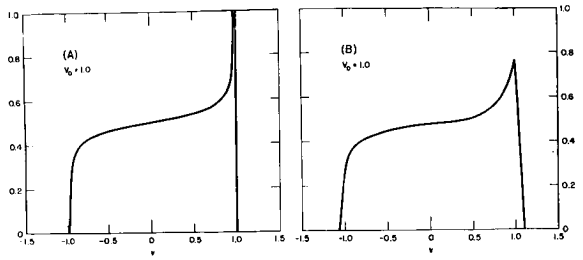


FIGURE 2.—Profiles of emission lines according to the proposed model. The emitting atoms have (A) a uniform speed of $V=1$ toward the secondary and (B) speeds of falling toward the secondary distributed according to Eq. (5) with $V_0=1$ and $n=1.1$.

pendent of φ, it follows that the intensity of emission line between v and $v+dv$ should be given by

$$I(v;\ V)\ dv = A\ dv \int d\varphi, \tag{1}$$

where A is a normalizing factor and the wavelengths are expressed in terms of velocity.

Since the inclination of the orbital plane, and consequently of the opaque disk, has been assumed to be different from 90°, we observe only one half of the emitting atoms, the other half being obscured by the opaque disk. Of course, if the radius of the opaque disk is not much larger than the radius of the sphere in which the emitting atoms are confined, we can still see part of the other half. However, in a binary system like β Lyrae, where the primary component is continuously and energetically ejecting mass, the equatorial plane must be populated with absorbing material even outside the opaque disk. Perhaps the equatorial plane outside the opaque disk is transparent to continuous radiation, but the opacity at wavelengths inside spectral lines is most likely appreciable. Therefore, it is reasonable to assume that only the emitting atoms located on this side of the obscuring disk impress their mark on the photographic plates. Thus the limits in the integral in equation (1) can be found easily in terms of θ or v. A simple geometrical consideration shows that for $0<\theta<\theta_1$, where $\theta_1=\frac{1}{2}\pi-i$, the integral extends from $\varphi=0$ to $\varphi=2\pi$. As θ becomes greater and greater, the range of the integral diminishes continuously and finally vanishes at $\theta=\frac{1}{2}\pi+i$. In this way we derive the broadening function as follows:

$$I(v;\ V)\ dv = dv/V \qquad (V\cos\theta_1 \le v \le V), \tag{2}$$

$$I(v;\ V)\ dv = (dv/\pi V)\left\{\pi - \cos^{-1}\left[\frac{v\tan\theta_1}{(V^2-v^2)^{1/2}}\right]\right\}$$

$$(0 \le v \le V\cos\theta_1), \tag{3}$$

$$I(v;\ V)\ dv = (dv/\pi V)\cos^{-1}\left[\frac{-v\tan\theta_1}{(V^2-v^2)^{1/2}}\right]$$

$$(-V\cos\theta_1 \le v \le 0). \tag{4}$$

Figure 2(A) illustrates the profile of this broadening function for $V=1$ and $\cos\theta_1=0.98$, which corresponds to an inclination of about 78°.5.

The broadening function just derived shows the basic feature of the emission lines observed in β Lyrae, namely, a positively shifted emission peak superimposed on the broad background emission (Sahade et al. 1959).

The previous calculation represents only an oversimplified model for the formation of emission lines. In reality, the emitting atoms do not fall into the secondary component with the same speed. The line profile given by equations (2)–(4) should be further broadened by the nonuniformity of falling speeds of emitting atoms.

Since the time for a particle of speed V to travel a distance ds is ds/v, the distribution function of V should be $1/V$. If the cutoff speeds at both ends are V_0 and nV_0, respectively, where n denotes a numerical factor greater than 1, the normalized distribution function of V will be

$$f(V)\ dV = (\ln n)^{-1}\ dV/V, \tag{5}$$

and the profile of emission will now be given by

$$I(v)\ dv = dv \int I(v,\ V) f(V)\ dV. \tag{6}$$

Evaluating the integral in equation (6) after equations (2)–(4) have been substituted, we obtain the final profile, which can be represented in six intervals of v by six different expressions.

We have computed the profile with $V_0=1$, $n=1.1$, and $\cos\theta_1=0.98$ according to the six expressions and have presented the result in figure 2(B). The profile shows a broad feature and a peak shifted toward the long wavelength, as observed (Houziaux, 1958). However, the computed peak is not so high as the observed one.

Since it would be unreasonable to assume that $n<1.1$ and since the predicted profile will have an even lower peak if $n>1.1$, we have to admit that the simple model of a spherically symmetric distribution of emitting atoms cannot quantitatively explain the observed emission peak, although it is adequate to account for the broad emission feature that moves with the secondary component.

To understand the emission peak, we should remember that the infalling emitting atoms must be more densely distributed near the plane of the disk. It can be seen easily that this asymmetry in distribution of emitting atoms tends to increase the intensity of the emission peak.

According to our model, the shift of the emission peak is always toward the long-wavelength side and depends on the velocity of infalling atoms. Since the shift persists in all phases, it reflects in the shift of γ velocity, as observed. The exact value of the radial velocity corresponding to the emission peak, however, depends on several factors: (1) the motion of the secondary component, (2) the velocity distribution of the falling atoms, and (3) the deviation from spherical symmetry. Consequently, its velocity variation may not be a good measure of the orbital motion of the secondary component. The radial velocity estimated from the center of the broad emission feature is not affected as sensitively by the second factor, as can be seen from figure 2(B); but it is difficult to measure its position in practice because of the absorption line or lines that cut into it.

That the emission lines move in phase with the secondary component suggests that the secondary component is much smaller than its corresponding lobe of the inner contact surface. Otherwise, there will be no space in which to put the emitting material. A natural deduction from this suggestion is that the secondary is smaller but more massive than the primary.

EVOLUTIONARY SIGNIFICANCE

Why should the primary component, which is less massive and therefore evolves more slowly, have reached the giant stage, while the more massive secondary component is underluminous? Sahade (1958; also see Struve, 1958) suggested that, as a result of more rapid evolution, the more massive secondary component had already passed the giant stage and had now assumed a position on the H–R diagram below the main sequence. In addition to Sahade's interpretation, we venture to propose here two more possibilities for the sake of further investigation.

1. The angular momentum of the prestellar material that finally condensed to become the secondary component may be extremely large and therefore retard the process of its contraction, as a star had to dissipate the angular momentum first before successful contraction. Consequently, the secondary component has, to begin with, a long lag in evolution. The rotating disk we have proposed, in order to explain various observational results, may represent the remnant of the prestellar gases of high angular momenta.

2. Originally the mass of the primary is larger than the secondary. When the primary has reached the giant or supergiant stage (e.g., Schwarzschild, 1958) and starts to eject mass through the inner contact surface, the secondary is still on the main sequence because of its relatively slow rate of evolution. The ejected mass of the primary falls into the secondary, as would be expected. Since the primary is more massive to begin with, a mass transfer in this way renders the separation between the two components smaller and smaller (Huang, 1963)—a process which, when combined with the effect of change in mass ratio of the two components, makes that lobe of the inner contact surface around the primary shrink fast and therefore further enhances the ejection of mass from the primary. For this reason, the time scale of a close binary in this mode of mass exchange is necessarily short; this explains why β Lyrae is an unusual object in the galaxy. The shortening of the separation will stop only when the masses of two components become equal. After this point, a further transfer of mass from the primary to the secondary will reverse the trend and widen the separation. This is the present situation of β Lyrae. Perhaps the absolute dimension of the lobe of inner contact surface around the primary does not change greatly at this stage, as the effect of increase in separation is compensated for by that of the change in mass ratio. Therefore, according to our interpretation,

the phase of drastic exchange of mass between the two components is now over. Soon (astronomically) there will be no transfer of mass and consequently no increase in period or widening of separation. Actually, this tendency is already shown incipiently in the equation of observed times of light-minima (e.g., Sahade et al. 1959), from which we can see that, while the third term definitely indicates the increase in period with time, the rate of increase is slowing down because of the negative sign in the fourth time.

This kind of mechanism of mass exchange between the two components of a close binary was first suggested by Crawford (1955) and by Kopal (1955) for explaining the characteristic properties of the Algol-type binaries, although Kopal (1959) has since reversed his stand. For whatever merit the mechanism has for explaining the Algol-type binaries, we consider that this mechanism is what is happening to β Lyrae right now.

Because of the mass accretion, the secondary is perhaps in a stage of gravitational contraction of the accreted mass and therefore is underluminous with respect to its present mass. This explains, in addition to the obscuring disk, why we do not observe light from the secondary even during principal eclipse. However, when its thermal state is adjusted to a level corresponding to its new mass, it will become a main-sequence star. Eventually it will evolve to the giant stage and perhaps eject mass. A transfer of mass from the secondary back to the primary will start a new cycle of change in the period, as well as in the separation, in the manner we have described.

Long before that, the secondary component would be more luminous, and the names of primary and secondary components would have to be interchanged in the astronomically not distant future. Therefore, in β Lyrae we obtain a situation in which the brighter component always fills the equipotential lobe and ejects mass toward the fainter component.

REFERENCES

Crawford, J. A., *Ap. J.*, **121**, 71, 1955.
Gaposchkin, S., *Zs. f. Ap.*, **39**, 133, 1956.
Guthnick, P., *Abh. Deutsch. Akad. Wiss. Berlin, Mathnaturw. Kla.*, No. 1, 1945–46.
Houziaux, L., *Pub. A.S.P.*, **70**, 209, 1958.
Huang, S.-S., *Ap. J.*, **136**, 903, 1962.
———, *Ap. J.*, **138**, 471, 1963.
Joy, A. H., *Pub. A.S.P.*, **54**, 35, 1942.
———, *ibid.*, **59**, 171, 1947.
Kopal, Z., *Mem. R. Sci. Liege*, **15**, 684, 1955.
———, *Close Binary Systems* (London: Chapman & Hall), 1959.
Kuiper, G. P., *Ap. J.*, **93**, 133, 1941.
Mitchell, R. I., *Ap. J.*, **120**, 274, 1954.
Sahade, J., *Ap. J.*, **63**, 52, 1958.
———, Huang, S.-S., Struve, O., and Zebergs, V., *Trans. Amer. Phil. Soc.*, **49**, 1, 1959.
Schwarzschild, M., *Structure and Evolution of the Stars* (Princeton University Press), 1958.
Stebbins, J., *Lick Obs. Bull.*, **8**, 186, 1916.
Struve, O., *Ap. J.*, **93**, 104, 1941.
———, *Stellar Evolution* (Princeton: Princeton University Press), 1950.
———, *Non-stable Stars*, ed. G. H. Herbig (Cambridge, Cambridge University Press), chap. 14, 1957.
———, *Pub. A.S.P.*, **70**, 5, 1958.
Wood, D. B., and Walker, M. F., *Ap. J.*, **131**, 363, 1960.
Woolf, N. J., *M.N.*, **123**, 399, 1962.

JEANS' CRITERION OF GRAVITATIONAL INSTABILITY

SU-SHU HUANG

Goddard Space Flight Center

Five years after Newton had published his law of gravitation, some one wrote him, asking whether the gravitational force would account for the condensation of matter into stars. In a letter dated December 10, 1692, Newton replied*:

"It seems to me, that if the matter of our sun and planets, and all the matter of the universe, were evenly scattered throughout all the heavens, and every particle had an innate gravity towards all the rest, and the whole space throughout which this matter was scattered, was finite, the matter on the outside of this space would by its gravity tend towards all the matter on the inside and by consequence fall down into the middle of the whole space, and there compose one great spherical mass. But if the matter were evenly disposed throughout an infinite space, it could never convene into one mass; but some of it would convene into one mass and some into another, so as to make an infinite number of great masses, scattered great distances from one to another throughout all that infinite space. And thus might the sun and fixed stars be formed, supposing the matter were of a lucid nature."

Thus, Newton, by his physical insight anticipated theory of condensation of matter in the universe more than two hundred years before Jeans in 1902 gave an exact mathematical treatment for the problem.

Jeans' treatment is a highly involved one and we cannot reproduce it here. However, we may describe the physical ideas underlying his mathematical calculations and then see whether we may derive his result without going too deeply into mathematics.

*This letter is quoted in Jeans' well known popular book, "The Universe Around Us," published by Cambridge University Press, 1944 (4th ed.).

Jeans considers an extended medium of gas and examines what would happen if a wave of disturbance should be created in this medium. Now let us first recall our common experience. When we throw a stone into a still pond of water, a train of waves is produced by this disturbance. When we strike a bell in a belfry, chime is heard miles away. The disturbance—this time the vibration of the bell—creates in the medium of air a series of sound waves which carry the vibration into a large distance.

The waves Jeans has considered in the extended medium is nothing more than a kind of sound waves. In physics we have learned that sound waves are composed of alternative condensations and rarefactions. A condensation holds together temporarily a large amount of molecules in the medium only to be dispersed again into a rarefaction as the wave passes. Because of the mass concentration the condensation exerts a higher gravitational attraction on the neighboring particles than anywhere in the medium outside condensations. In the ordinary sound waves this excess in gravitational pull of a condensation is infinitesimal and has absolutely no perceptible effect. But when the same phenomenon happens on a large scale, the gravitational force of the condensations may become very important.

Now waves are measured by the wavelength. In the case of sound waves, the wavelength is the range covered by a condensation and a rarefaction, which correspond respectively to a crest and a trough in water waves. Hence the size of the condensation increases with the wavelength. And as the size of the condensation increases indefinitely, a state will be reached such that the mutual gravitational attractions among the particles in the temporary condensation become

so strong that they will not disperse afterwards. The condensations become permanent. When this happens, the extended medium is said to be gravitationally instable because it will not remain as an extended medium. Instead it will coagulate into many big lumps of masses, as Newton has envisaged.

From a long mathematical analysis Jeans derived that the critical wavelength, λ_0, of density fluctuation that will cause instability in an extended medium is given by

$$\lambda_0^2 = \pi\gamma V^2/3G\rho$$

which is Jeans' criterion of gravitational instability; any disturbance that has a wavelength equal to or greater than λ_0 produces a collapse of the extended medium into individual permanent condensations. In the equation, ρ is the density of the medium, V is the mean velocity of the particles of which the medium is formed, γ is the ratio of specific heats which also depends upon the nature of gas in the medium, and G is the usual gravitational constant. The gas which pervades interstellar space is dominantly composed of atomic hydrogen, so γ may be set equal to 5/3.

The radius of a condensation is roughly equal to one-half of the wavelength because, as we have said, a condensation and a rarefaction together stretch a complete wavelength. Therefore, according to Jeans' criterion, the radius of permanent condensation, R, should satisfy the following inequality,

$$R^2 \geq 0.44\ V^2/G\rho \quad \text{in c.g.s. units.}$$

Before we begin to discuss the consequences of gravitational instability in an extended medium, let us now examine whether we may obtain Jeans' formula by some simple calculations. Indeed if we look at the physical situation that leads to condensations, we do find a simple way for deriving Jeans' criterion.

Consider a spherical condensation of mass M and radius R. What is the minimum velocity that a particle on its surface must have in order to escape from the condensation? This minimum velocity is obviously what is known as the escape velocity and is given by the well known expression $(2GM/R)^{1/2}$. According to this expression the escape velocity of the earth for example is 11 km/sec while that of the moon is 2.4 km/sec. This explains why the moon cannot hold an atmosphere. The velocities due to molecular motion at temperatures existing on the surface of the moon would be greater than its escape velocity and all molecules would escape it even if it originally had an atmosphere.

Similarly, for the condensation we have discussed, the molecular velocity V must be smaller than its escape velocity, if the condensation is going to hold itself permanently. Otherwise the entire condensation will evaporate away because the condensation, unlike the solid body of the moon, is composed completely of gaseous and dust particles. Therefore from the present simple consideration, the criterion of instability in the extended medium should be

$$V^2 \leq 2GM/R$$

which is equivalent to

$$R^2 \geq 0.12\ V^2/G\rho \quad \text{in c.g.s. units,}$$

if we remember that $M = (4\pi/3)R^3\rho$. When we compare the present inequality with Jeans' original one, we have admitted that we did not obtain exactly the same expression. However, the two differ only in the numerical factor. The radius itself derived in the present way is about one half that given by Jeans. However, we should not be overconcerned with this difference in the numerical factor. In problems such as the condensation in an extended medium we only expect an answer in the order of magnitude and a numerical factor of 2 does not make much difference to our understanding of the problem. On the other hand this simple derivation gives us a clear concept of why an extended medium should become instable.

We can now obtain an expression of the total mass, M, of the condensation directly from Jeans' original formula if we again make use of the relation $M = 4\pi R^3\rho/3$. Instead of density, ρ, astronomers often measure the mass concentration in the interstellar medium in terms of the number, N, of hydrogen atoms per cm^3. This means that the mass of N hydrogen atoms per cm^3 is equal to the density but does not imply that all matter in the medium is in the form of hydrogen atoms,

although they are indeed the dominant constituent. Expressing V now in km/sec we obtain

$$M = 2.73 \times 10^4 \, V^3/N^{1/2} \text{ solar mass}$$

after the numerical factor has been evaluated. It is interesting to note that M is inversely proportional to $N^{1/2}$, because in the medium of a small density it needs a large mass to hold the constituent particles together permanently.

The density in interstellar media fluctuates greatly from one region to another, ranging from 1 to well over 1000 hydrogen atoms per cm^3. There is evidence indicating that interstellar matter is in a state of turbulent motion. In fact such a turbulent motion may be regarded as the cause of disturbance which leads to density waves in the medium just as the wind in the atmosphere produces sound waves.

It is difficult to estimate the turbulent velocities in the medium. On the other hand thermal velocities of hydrogen atoms in the regions remote from hot and luminous stars (i.e., in the so-called HI region) have been estimated to be 1–2 km/sec which corresponds roughly a temperature between 10° to 100°K. In the accompanying table we have listed the lower limit of M for several possible combinations of V and N. From the reasoning given before, V actually should include turbulent velocity of the medium and will assume a much higher value than what are given in the table. Also, in the interstellar space near hot and luminous stars where hydrogen atoms are ionized (i.e., in the so-called HII region) thermal velocities of hydrogen atoms are much higher. Consequently the masses of condensations in the interstellar space could be much greater than the tabulated lower limit.

TABLE 1.—*Masses of Condensations in $M\odot$ According to Jeans' Criterion of Instability*

N in No. of H/cm^3	V in km/sec	
	1	2
1	2.73×10^4	2.18×10^5
10	8.63×10^3	6.91×10^4
100	2.73×10^3	2.18×10^4
1000	8.63×10^2	6.91×10^3

A close examination of the table leads us to conclude that the first fragmentation of the gaseous medium that was to become our Milky Way system cannot be stars because the mass of each condensation is too large. Considering the huge amount of mass in each condensation we would suggest that it represents a mass conglomeration which evolves to become a star cluster.

Perhaps the globular cluster were first fragmented in this way from the general galactic medium when the latter was still confined in a spherical volume. The condensation resulting from the first fragmentation will contract because of the mutual gravitational forces among its parts. As the condensation contracts, the density increases. A second fragmentation following again Jeans' mode of instability may occur in the contracting condensation. Because of a large density in it, the masses resulting from the second condensation process will be small according to the formula given before and may be of the order of stellar masses. In this way stars are formed in the clusters. However, astronomers' opinions are divided as to whether the stars are indeed fragmented according to Jeans' mode in the second process.

While stars are being shaped in the globular cluster, the molecules in the galactic mass suffer numerous collisions themselves. Now if the matter does not possess a net angular momentum, such collisions would not affect the general shape of the body. However, if there exists a net angular momentum as does the primaeval galaxy, the collisions would tend to modify gradually the general form of the galaxy from a spherical to a disk-like structure. This is due to the dissipation of kinetic energy in the direction perpendicular to the plane of rotation associated with the net angular momentum. It can be easily seen if we consider two streams of particles colliding in that direction. Their velocities will neutralize each other and the kinetic energy will be dissipated away through radiation. But if the collisions take place in the plane of rotation, the velocities of particles cannot completely cancel each other because the angular momentum has to be con-

served. In this way the vertical velocity components of particles decrease continuously. As a result the gaseous condensation becomes flattened into a disk which is the galactic plane. Thus, objects in the galactic plane still move rapidly around the center of the galaxy. In the solar neighborhood this rotational velocity is of the order of 200 km/sec, but the vertical components of their velocities are small compared to this value. Globular clusters and stars already formed when the primaeval galaxy was still of a spherical form would not collapse into the disk because they do not collide frequently among themselves as the gaseous and dust particles do. Unless they collide bodily kinetic energy cannot be dissipated through gravitational interaction alone, provided that we do not consider the dissipation through gravitational waves which is exceedingly small anyway if present at all. Hence the globular clusters and other extremely population II objects maintain a substratum of spherical distribution in the galaxy to this day.

Once the gaseous and dust particles have collapsed into the galactic plane, the medium will also become unstable according to Jeans' criterion. Perhaps the condensation thus formed are those interstellar clouds of which we now see plenty in the galactic plane. Stars are then formed in these clouds following a process which may or may not be related to Jeans' mode of instability. In any case the galactic clusters and associations are formed in the clouds. A nearby cloud complex exists in the Orion region, which is rich in hot and luminous stars of spectral types O and B. Some of the most conspicuous nebulae in the sky including the famous emission nebula in Orion belong to this complex.

That O and B stars are continually being formed in the cloud is now beyond dispute because the O and early B stars with a life time less than 10^8 years on the main sequence must be newly formed. However, whether their formation is a direct consequence of Jeans' gravitational instability is not certain.

Many astronomers including Jeans himself have also suggested that galaxies themselves are formed from fragmentation of the primaeval medium in space according to Jeans' mode of instability. This is of course precisely what was conjectured by Newton in the letter we have quoted in the beginning of this article. However, how the galaxies were formed depend greatly on the structure (or model) of the universe. Also, an application of Jeans' criterion to the fragmentation of galaxies from the primaeval medium that pervaded everywhere in the universe is complicated by the fact that the universe is expanding. Therefore the importance of Jeans' gravitational instability in its astrophysical applications may not be as ubiquitous as when Jeans first formulated it. But it remains a fundamental conception that every astronomer should remember.

MODES OF MASS EJECTION BY BINARY STARS AND THE EFFECT ON THEIR ORBITAL PERIODS

SU-SHU HUANG

Goddard Space Flight Center

Three modes of mass ejection from one or both components of a binary system have been idealized and their physical characteristics discussed, in order to predict their effect on the orbital period of the system. Also, the consequence of synchronization of axial rotation of component stars with their orbital revolution is examined.

I. INTRODUCTION

The observed period of an eclipsing binary is defined as the time interval between two consecutive minima of light-variation. Thus, even if the intrinsic period of the binary itself remains invariable, the observed period can still undergo periodic changes when other orbital elements, such as the longitude of periastron, vary with time (Tisserand, 1895; Kopal, 1959). The complicated variations in the observed periods of many eclipsing binaries may be due to this mechanism. On the other hand, Wood (1950) has suggested that the change in orbital period may also be caused by ejection of mass. Indeed, if the direction of ejection is at random, like the prominence activities in the solar atmosphere, the mass ejection over a long period of time would statistically follow a spherically symmetric pattern. However, in a relatively short interval the ejection of mass may be more pronounced in the direction of orbital motion of the component star than in the reverse direction at one time and vice versa at another time, resulting in a complicated change in the orbital period. For this reason, Wood's suggestion merits further study in our search for the underlying reason for the change in orbital periods of many eclipsing binaries, although this opinion is not shared by some (e.g., Kopal, 1959).

Our interest in the mass variation of stars in a binary system arises from the study of β Lyrae. Its period has been increasing since the earliest record (Kopal, 1941), indicating no periodicity in the variation. This result, combined with the evidence derived from other empirical data (Kuiper, 1941; Struve, 1941), led us to suggest that the increase in period may be attributed to the mass loss by one or both of the two components stars (Huang, 1956). More recently, because of the change in our understanding as regards the relative masses of the two components of this peculiar system (Huang, 1962a; Woolf, 1962a), the increase in period may also be attributed to the mass transfer from the less massive but more luminous primary to the more massive but less luminous secondary component (Woolf, 1962b).

The effect of mass ejection by one or both of the components or of mass transfer from one to the other on the orbit of the binary depends not only upon the amount of mass lost or transferred but also upon the mode in which the mass is being lost or transferred. We propose here to formulate a few idealized cases of the mass-loss process and examine the effect on orbital period in each idealized case.

The variations in period, P, in semimajor axis, a, and in eccentricity, e, of the relative orbit in terms of the variations in masses, m_1 and m_2, of the two components, in orbital angular momentum per unit reduced mass of the system, h, and in

the dynamical energy of the system per unit reduced mass, E, can be written as follows (Huang, 1956):

$$\frac{\delta a}{a}=\frac{\delta(m_1+m_2)}{m_1+m_2}-\frac{\delta E}{E}, \tag{1}$$

$$\frac{\delta P}{P}=\frac{\delta(m_1+m_2)}{m_1+m_2}-\frac{3}{2}\frac{\delta E}{E}, \tag{2}$$

and

$$\frac{e\ \delta e}{1-e^2}=\frac{\delta(m_1+m_2)}{m_1+m_2}-\frac{1}{2}\frac{\delta E}{E}-\frac{\delta h}{h}. \tag{3}$$

Equations (1)–(3) are formal in nature because they express the changes in orbital elements in terms of the changes in E and h, both of which depend critically on the mode of mass ejection or mass transfer and are difficult to evaluate in general except in the few idealized cases that will be discussed below.

II. JEANS'S MODE

One of these idealized cases was first studied by Jeans (1924, 1925) and will hereafter be called "Jeans's mode of mass ejection." It assumes that the mass of the component star decreases gradually, leaving no reaction on the binary system by this process. This assumption is realizable in the loss of mass of a star in the form of radiation as a result of the equivalence of energy and mass. Indeed, this is what prompted Jeans to study this mode. Since then it has been applied to cases involving direct loss of mass (Huang, 1956; Boersma, 1961). The physical circumstance of actual mass ejection that may be approximated by Jeans's mode must satisfy two conditions: (1) the ejection of mass has a satistically spherical symmetry and (2) the velocities of ejection must be very high, for otherwise the ejected mass would exert a strong action on the star and consequently depart from the underlying assumption of the mode. According to these conditions, the loss of mass resulting from a supernova explosion, as proposed by Blaauw (1961) for the origin of the O- and B-type stars with high velocities, and the loss of mass by the ordinary novae (Ahnert, 1959) would closely approximate this mode.

It can easily be shown (Huang, 1956) that Jeans's mode of ejection leads to

$$\delta h/h=0$$

and

$$\delta E/E=2\delta(m_1+m_2)/(m_1+m_2), \tag{4}$$

which immediately yields the following results due originally to Jeans:

$$\delta a/a=-\delta(m_1+m_2)/(m_1+m_2),$$

$$\delta P/P=-2\delta(m_1+m_2)/(m_1+m_2),$$

and

$$e\ \delta e/(1-e^2)=0, \tag{5}$$

when we substitute equations (4) in (1)–(3).

The parameters E and h, while convenient to use under some circumstances, are not simple physical quantities. We are more interested in the dynamical energy and the angular momentum per unit mass, denoted, respectively, by E_0 and h_0, than those per unit reduced mass E and h. Since E_0 and E are simply related by

$$E_0=[m_1m_2/(m_1+m_2)^2]E \tag{6}$$

with a similar relation between h_0 and h, we can readily obtain the variations in E_0 and h_0 in Jeans's mode of ejection as follows:

$$\frac{\delta h_0}{h_0}=\frac{\delta m_1}{m_1}+\frac{\delta m_2}{m_2}-\frac{2\delta(m_1+m_2)}{m_1+m_2}, \tag{7}$$

$$\frac{\delta E_0}{E_0}=\frac{\delta m_1}{m_1}+\frac{\delta m_2}{m_2}. \tag{8}$$

Thus, in contradistinction to h, the angular momentum per unit mass changes in Jeans's mode.

If only one component is ejecting mass, we may set $\delta m_2=0$. It follows from equation (7) that $\delta h_0>0$ if $m_1>m_2$ and $\delta h_0<0$ if $m_1<m_2$ because δm_1 is negative. This result is self-evident, as the ejection from the more massive component carries away an angular momentum per unit mass less than the average value and thus leaves the remaining system a higher h_0 than before.

In equation (8) we may note the dynamical energy per unit mass of the system has increased as a result of mass loss in Jeans's mode, since negative δm_1, δm_2, and E_0 necessarily lead to a positive δE_0. Thus not only is energy supplied to the ejected matter, but also the dynamical energy

per unit mass of the binary system has increased as the result of mass ejection.

An extreme case of Jeans's mode of ejecting matter from a binary star is that in which the ejection takes place instantaneously rather than continuously. If the ejected mass exceeds a certain amount, it would result in a complete separation of the two component stars, as was pointed out by Blaauw (1961), who has assumed a circular orbit for the binary. A generalization to elliptical orbit is straightforward and yields the result that the two components of a binary system will become two single stars if the amount of ejected mass, $-\Delta(m_1+m_2)$, satisfies the following condition:

$$-\Delta(m_1+m_2)\geq\tfrac{1}{2}(r/a)(m_1+m_2), \qquad (9)$$

where r represents the separation of the two stars at the instant when the instantaneous mass ejection, according to Jeans' mode, takes place. It follows from relation (9) that, in order to dissociate a binary, it is most favorable to have mass ejection occur at or near the periastron and least favorable at or near the apastron. Now when the orbit is eccentric, the components spend most of their time near the apastron. Therefore, we would expect binaries with high eccentricities to be less likely to be dissociated in this way.

III. SLOW MODE

While Jeans's mode represents one extreme of mass loss with high velocities, the other extreme of losing mass with slow velocities provide us with the second idealized case that is tractable. When the ejection velocities from the component stars are low, the ejected matter will not be expected to escape out of the binary system. Consequently, the total angular momentum of the system will be conserved, although the total dynamical energy of the system can very either way, depending on the manner in which the ejected matter flows in in the system.

We may divide this mode into two cases: (1) the mass ejected is falling back either to the original star or to its compansion, and (2) the particles ejected from the less massive component form, after many collisions among themselves, a rotating ring around the more massive component. Perhaps the second case is observationally more important than the first because we cannot directly observe the mass transfer from one to the other, but many Algol-type variables are actually found to have rotating rings around their more massive components (Joy, 1942, 1947; also Sahade, 1960).

Since we do not know the variation in E_0 in both cases of the slow mode of ejection, we cannot determine $\delta P/P$ (or $\delta a/a$) and $e\,\delta e/(1-e^2)$ simultaneously. However, this does not prevent us from expressing $\delta P/P$ in terms of $\delta h/h$ and $e\,\delta e/(1-e^2)$. By eliminating $\delta E/E$ from equations (2) and (3) and expressing $\delta h/h$ in terms of $\delta h_0/h_0$, we obtain the following equation for the change in orbital period:

$$\frac{\delta P}{P}=\frac{4\delta(m_1+m_2)}{m_1+m_2}-3\left(\frac{\delta m_1}{m_1}+\frac{\delta m_2}{m_2}\right)+3\frac{\delta h_0}{h_0}+\frac{3e\ \delta e}{1-e^2} \qquad (10)$$

and a corresponding equation for $\delta a/a$.

The equations for $\delta P/P$ and $\delta a/a$ thus obtained are quite general. Now, in the case of simple mass transfer from one component to another, the total mass and the total angular momentum of the system must be conserved. Hence we have

$$\delta(m_1+m_2)=0 \quad \text{and} \quad \delta h_0=0, \qquad (11)$$

which, when substituted in equation (10), leads to

$$\frac{\delta P}{P}=-3\left(\frac{m_2-m_1}{m_2}\right)\frac{\delta m_1}{m_1}+\frac{3e\ \delta e}{1-e^2}. \qquad (12)$$

For binaries of small eccentricities, the second term in the right-hand side of equation (12) may be neglected. Thus we arrive at the result that a transfer of mass from the more massive component to the less massive component (i.e., $m_1>m_2$ and $\delta m_1<0$) makes the period decrease with time, and a transfer of mass in the reverse direction (i.e., $m_2>m_1$, $\delta m_1<0$) results in an increasing period.

In explaining the peculiar behavior of the β Lyrae system, we have suggested that the primary, which is less massive now, was originally more massive (Huang, 1962b). The reversal of relative massiveness between the two components can be most efficiently realized by a direct transfer from the primary to the secondary component. That the period of β Lyrae increases constantly

provides independent evidence that its primary component that ejects gases is the less massive one.

In accordanse with previous results (Huang, 1962a; Woolf, 1962a), we assume the mass of the primary of β Lyrae to be $5m\odot$ and that of the secondary $15m\odot$. If we use the average value of dP/dt of 9.38 seconds per year as the rate of increase in period (e.g., Kopal, 1941), we find from equation (12) that the rate of transfer of mass from the primary to the secondary must be

$$dm_1/dt = -2.1\times10^{-5}\ m\odot/\text{year}, \qquad (13)$$

in order to account for the observed rate of increase in period. That the increase in period of β Lyrae is due to mass transfer from the primary to the secondary was first mentioned by Woolf (1962b).

On the other hand, if the primary m_1 is ejecting mass according to Jeans's mode, we have

$$dm_1/dt = -8.4\times10^{-5}\ m\odot/\text{year}, \qquad (14)$$

which is equivalent to 5.3×10^{21} gm/sec. From the data obtainable from the spectroscopic observation, Struve (1958) has shown that the rate of mass loss of the system cannot be considerably greater than 4×10^{20} gm/sec. Unless the mass is dissipated mainly through corpuscular radiation—a possibility that cannot be entirely ruled out—we must again accept the conclusion that mass is being transferred between the two components of this peculiar system.

The second possibility of the slow mode of ejection is the formation of a rotating ring around the more massive component by the matter ejected from the less massive component. Qualitative arguments can be given to show that the formation of a ring around the less massive component due to ejection of the massive component is unlikely (Huang, 1957).

The change in orbital elements as a result of ring formation may be easily formulated under some simplifying assumptions. Consequently, we may suppose that, in the process of ring formation around the more massive component, there is a change only in orbital angular momentum but no change in the total mass of the system, namely, $\delta(m_1+m_2)=0$. If we assume, without loss of generality, that the less massive m_2 is ejecting mass to form a ring rotating around m_1 in the same sense as the binary motion, we have

$$\delta h_0/h_0 = \gamma_1(\delta m_2/m_2); \qquad (15)$$

where

$$\gamma_1 = [(m_1+m_2)a_i/m_1a(1-e^2)]^{1/2} \qquad (16)$$

and a_i is the radius of the ring around m_1. In deriving equation (15), we have assumed that the circular motion of the ring is under the gravitational attraction of m_1 alone. It follows from equations (10) and (15) that

$$\frac{\delta P}{P} = \alpha\frac{\delta m_2}{m_2} + \frac{3e\delta e}{1-e^2}, \qquad (17)$$

where

$$\alpha = 3\left(\gamma_1 - 1 + \frac{m_2}{m_1}\right). \qquad (18)$$

TABLE 1.—*Values of α for Different Combinations of a_i/a and m_2/m_1 ($e=0$)*

a_i/a	α				
	$m_2/m_1=\frac{1}{2}$	$m_2/m_1=\frac{1}{3}$	$m_2/m_1=\frac{1}{4}$	$m_2/m_1=\frac{1}{8}$	$m_2/m_1=\frac{1}{10}$
0.2	+0.14	−0.45	−0.75	−1.05	−1.29
0.3	+0.51	−0.10	−0.41	−0.73	−0.98
0.4	+0.82	−0.19	−0.13	−0.45	−0.71
0.5	+0.10	+0.45	+0.12	−0.21	−0.48
0.6	+1.35	+0.68	+0.35	+0.01	−0.26
0.7	+1.57	+0.90	+0.56	+0.21	−0.07
0.8	+1.79	+1.10	+0.75	+0.40	+0.11

The value of a_i that enters into γ_1, according to equation (16), depends on the amount of angular momentum that the ejected matter carries away from m_2, i.e., on the mode of ejection.

The value of α depends on the relative size of the ring as compared with the semi-major axis axis of the binary system, as well as the mass ratio of the two components. Table 1 lists values of α for a few combinations of a_i/a and m_2/m_1 for the case $e=0$. Those cases represented in the lower left corner are not realistic, as the ejected particles cannot form a stable ring under the strong perturbation by m_2 even if they possess the necessary angular momentum for forming it. The ejected matter will either escape or form a ring around the entire system—a case that will be discussed in the next section.

After we have excluded the cases entered in the lower left corner of the table, we are still left with values of α both positive and negative. Thus formation of a ring around the more massive component can change the orbital period in both ways, but the change will, in general, be small.

IV. INTERMEDIATE MODE

By "intermediate mode" of ejecting mass from a component of the binary system we mean that the ejection velocities are large enough to overcome the attraction of both components, so that ejected particles are able to penetrate the inner contact surface of the system (e.g., Kuiper, 1941). As a result, the mass may either escape or form a a ring rotating around the entire system. In the first case we may express the change in orbital period in terms of the ejected mass and the angular momentum carried away with it. If h_e is the angular momentum per unit mass that has escaped from the system, we have

$$\frac{\delta h_0}{h_0}=\frac{h_e-h_0}{h_0}\,\frac{\delta(m_1+m_2)}{m_1+m_2}. \tag{19}$$

Now, if $h_e>h_0$, i.e., if the escaped matter, $-\delta(m_1+m_2)$, carries away more than its average share of angular momentum, then δh_0 will be negative, as indeed one can see from equation (19), since $\delta(m_1+m_2)$ is negative. On the other hand, if $h_e<h_0$, δh_0 will be positive.

Substitution of equation (19) in equation (10) will give the required formula for the change in orbital period in terms of h_e. However, the equation thus derived is only of a formal nature, since we cannot estimate the value of h_e by any observational means.

The situation becomes more tractable if the escaped particles form an observable ring around the entire system. The ring, if it should be formed at all, must be large compared with the binary system itself, for otherwise it will not be stable under the attraction of two revolving stars. Hence, as a first approximation, the particles in the ring may be regarded as moving under the gravitational attraction of a mass m_1+m_2 located at the center of mass of the system. Furthermore, because of frequent collisions among particles themselves, we may expect the ring to be circular.

If the ring is revolving around the center of mass with radius, a_e, and in the same sense as the binary motion, we obtain, for the angular momentum per unit mass of the ring,

$$h_e=[G(m_1+m_2)a_e]^{1/2}, \tag{20}$$

which must come from the binary system itself. Now, if the mass in the ring is $-\delta(m_1+m_2)$, we derive from equations (10), (19), and (20) the following equation for the change in period:

$$\frac{\delta P}{P}=(1+3\gamma_2)\frac{\delta(m_1+m_2)}{m_1+m_2}-3\left(\frac{\delta m_1}{m_1}+\frac{\delta m_2}{m_2}\right)+\frac{3e\ \delta e}{1-e^2}, \tag{21}$$

where

$$\gamma_2=\frac{(m_1+m_2)^2}{m_1m_2}\left[\frac{a_e}{a(1-e^2)}\right]^{1/2}. \tag{22}$$

For binaries in circular motion, i.e., $e=0$, it can be shown that $\delta P/P$ decreases if

$$(a_e/a)\geq 1, \tag{23}$$

which is, of course, true for all rings formed around the entire system.

We should emphasize that the three modes of mass ejection are introduced only as idealized cases. The actual mode of mass ejection is most likely a combination of these idealized modes, since particles ejected from the component star are expected to have wide ranges in velocity and in angular momentum per unit mass. Even with frequent collisions among themselves, it is hard to imagine that they will completely equalize

their velocities and thereby follow any one of the three modes discussed previously. Thus, in the case of β Lyrae, we have concluded that the increase in its orbital period is most likely due to a transfer of mass from the primary to the secondary component. However, it should be noted that it also develops a ring, or a shell, around the entire system, as has been pointed out by Struve (1941). Consequently, the β Lyrae system is losing mass in at least two different modes. However, since the period is increasing, the dominant mode of mass ejection in the system probably follows the slow mode.

V. COUPLING BETWEEN ORBITAL MOTION AND AXIAL ROTATION

The finite size of component stars in a close binary system gives rise to a number of dynamical problems, such as the precession and nutation of the components (Brouwer, 1946). In the present section we shall consider only the problem of coupling between orbital motion and axial rotation in connection with mass ejection from the component stars, since angular momentum may be transferred from orbital motion to axial rotation, and vice versa, through tidal interaction. Complicated as the problem is quantitatively, we can nevertheless understand it in a more or less general way.

If the mass ejection occurs in a relatively short time compared with the time scale over which the tidal force produces an appreciable effect, the orbital and rotational angular momenta may be regarded as completely independent without any kind of coupling. Consequently, all results derived in the previous sections apply to this case. On the other hand, if mass ejection from the component star takes place over a period of time long enough for the orbital motion and rotation to be strongly coupled, we have the other extreme case in which the two kinds of motion are perpetually synchronized. Indeed, in most close binaries so far observed, synchronization appears to be the general rule (Swings, 1936; also Struve, 1950).

If axial rotation and orbital motion are synchronized, we may expect, as we shall assume in the following discussion, that the rotational axes of both components are perpendicular to the orbital plane. With this assumption we can proceed to study the change in orbital period as a result of synchronization.

We shall discuss it in the same sequence as we have studied the non-interacting cases in the previous three sections. However, Jeans's mode, which imposes the condition $\delta h = 0$, automatically fixes the amount of change in orbital angular momentum in the process and consequently precludes any consideration of axial rotation. Hence we have only to examine the remaining two modes.

In the first case of the slow mode, both total mass and total angular momentum, Ω, of the binary system are conserved. Since the latter is is given by

$$\Omega = (m_1+m_2)h_0 + (2\pi/P)(m_1k_1^2+m_2k_2^2), \quad (24)$$

where k_1 and k_2 are radii of gyration of the two component stars, we have, in the case of mass transfer from one to the other component, a change in h_0, due to the changes in P, m_1, m_2, k_1, and k_2, given by

$$\delta h_0/h_0 = (A_1+A_2)(\delta P/P) - \left(A_1\frac{\delta m_1}{m_1}+A_2\frac{\delta m_2}{m_2}+2A_1\frac{\delta k_1}{k_1}+2A_2\frac{\delta k_2}{k_2}\right), \quad (25)$$

where

$$A_1 = [(m_1+m_2)/m_2(1-e^2)^{1/2}](k_1/a)^2 \quad (26)$$

and

$$A_2 = [(m_1+m_2)/m_1(1-e^2)^{1/2}](k_2/a)^2. \quad (27)$$

The changes in k_1 and k_2 may be affected both by the process of mass transfer and by internal evolution of the star, such as in β Lyrae (e.g., Kopal, 1959).

Equation (24), when combined with equation (10), leads to the following result:

$$(1-3A_1-3A_2)(\delta P/P) = 3\left(\frac{1+A_2}{m_2}-\frac{1+A_1}{m_1}\right)\delta m_1+\frac{3e\ \delta e}{1-e^2} - 6\left(A_1\frac{\delta k_1}{k_1}+A_2\frac{\delta k_2}{k_2}\right). \quad (28)$$

If we compare this equation with the corresponding one in equation (12), we find that the difference arises only from terms involving A_1 and A_2. First, we have the multiplying factor

$(1-3A_1-3A_2)$, whose effect is to increase the variation in period as compared with the non-coupling case for the same amount of mass transferred from one component to the other. This is understandable, as a slight increase in P (or a) makes rotation correspondingly slower as a result of synchronization. Consequently, the angular momentum is fed from rotation to orbital motion to make the change in P greater than that predicted in the non-coupling case.

Second, the term, $3[(A_2/m_2)-(A_1/m_1)]\delta m_1$, which may be expressed in the form

$$\frac{3(m_1+m_2)}{m_1m_2(1-e^2)^{1/2}}[(k_2/a)^2-(k_1/a)^2]\delta m_1,$$

results from the change in rotational angular momentum as a result of transfer of mass. Thus, if, let us say, $k_1>k_2$ and if mass is transferred from m_1 to m_2, the condition of synchronization cuts down the rotational angular momentum in favor of orbital angular momentum. The net result is to increase the period more than would be expected from the noncoupling case, as is indeed shown by this positive term.

It is simple to interpret the terms involving δk_1 and δk_2, since an increase in k_1, k_2, or both will take angular momentum away from orbital motion, in order to keep up the state of synchronization and thereby reduce the period.

As regards the second case of slow mode in which a rotating ring is formed around the more massive component, we can similarly obtain the following equation for determining $\delta P/P$:

$$(1-3A_1-3A_2)(\delta P/P)$$
$$=\left(\frac{\alpha-3A_2}{m_2}+\frac{3A_1}{m_1}\right)\delta m_2+\frac{3e\ \delta e}{1-e^2}-6\left(A_1\frac{\delta k_1}{k_1}+A_2\frac{\delta k_2}{k_2}\right), \quad (29)$$

which corresponds to equation (17) in the non-interacting case.

In the intermediate mode, we have the variation in h_0 given by an equation similar to equation (19) but which now involves A_1 and A_2. This equation, together with equation (10), leads to an equation for $\delta P/P$ which involves, among other factors, h_e, i.e., the angular momentum per unit of escaped mass. If a rotating ring is formed around the entire system, so that h_e may be estimated, as in the previous section, the variation in P may be expressed in the following equation:

$$(1-3A_1-3A_2)(\delta P/P)=(1+3\gamma_2)$$
$$\times\frac{\delta(m_1+m_2)}{m_1+m_2}+\frac{3e\ \delta e}{1-e^2}-3(1+A_1)(\delta m_1/m_1)$$
$$-3(1+A_2)(\delta m_2/m_2)-6\left(A_1\frac{\delta k_1}{k_1}+A_2\frac{\delta k_2}{k_2}\right), \quad (30)$$

which corresponds to equation (21) in the uncoupled case.

Comparing the case of perpetual synchronization with the case of no coupling [i.e., eqs. (28), (29), and (30) versus, respectively, eqs. (12), (17), and (21)], we find that the difference between the two extremes of coupling in all cases lies only in two factors: one is the multiplying factor $(1-3A_1-3A_2)$ and the other is an additive term,

$$-3[A_1(\delta m_1/m_1)+A_2(\delta m_2/m_2)$$
$$+2A_1(\delta k_1/k_1)+2A_2(\delta k_2/k_2)].$$

Thus the seriousness of the difference may be seen from the magnitude of A_1 and A_2.

Let R_1 and R_2 be the radii of the two components. Then the ratio k_1/R_1 or k_2/R_2 may be calculated, once the stellar model is assumed. At the same time, the ratios R_1/a and R_2/a may be obtained from photometric data. Therefore, $(k_1/a)^2$ and $(k_2/a)^2$ can be estimated. If the mass ratio of the binary is also known, we can compute A_1 and A_2. Let us now examine the values of A_1 and A_2 for a few well-known binaries.

Without doubt the most important case of synchronization is the group of binaries of W Ursae Majoris type, because they are probably in physical contact and must be synchronized in rotation and orbital motion (e.g., Struve, 1950). For this type of star we may roughly set $m_1=m_2$ and use the values of $(k/R)^2=0.065$ computed from a solar model (Motz, 1953), for both components. The value of R/a may be set to $\frac{1}{2}$ for both components because they are in contact. In this way we find that A_1 and A_2 are approximately equal to 0.025. Hence the correction factor due to synchronization is quite appreciable.

In the case of the Algol-type binaries, the difference in $\delta P/P$ as predicted from the two extreme

cases of coupling is slightly smaller. In the case of Algol itself, we have $R_1/a=0.23$ and $R_2/a=0.24$. If we take the solar value for $(k_1/R_1)^2$ of the main-sequence primary component and a value of 0.012 (Motz, 1952) corresponding to a giant for $(k_2/R_2)^2$ of the subgiant secondary component, we have $A_1=0.021$ and $A_2=0.0007$ if we adopt a mass ratio $m_1/m_3=5$ (Meltzer, 1957). Therefore, for Algol and, in general, for systems like it, A_2 may be neglected. However, A_1 is large enough to make the results in the coupled case significantly different from those in the uncoupled one.

If axial rotation and orbital motion are neither completely coupled nor completely independent, the problem becomes much more difficult because then a satisfactory treatment must take into account the time scale of mass ejection, as well as that of momentum transfer between the two kinds of motion. Formally, however, we may treat the problem by assuming that only a part of the angular momentum of axial rotation is coupled to orbital motion and use the following interpolation formula:

$$(m_1+m_2)h_0+(2\pi c/P)(m_1k_1^2+m_2k_2^2)$$

in place of the total angular momentum given by equation (24). Here c is a parameter that is connected with the degree of coupling. Thus $c=0$ for the case of no coupling (secs. II–IV) and $c=1$ for the case of synchronization. Intermediate coupling may be expected to give values of c between these two limits. The variation in orbital period may now be simply obtained by multiplying every A_1 and A_2 by c in the resulting equations for $\delta P/P$.

REFERENCES

AHNERT, P., *A.N.*, **285,** 191 (1959).
BLAAUW, A., *B.A.N.*, **15,** 265 (1961).
BOERSMA, J., *B.A.N.*, **15,** 291 (1961).
BROUWER, D., *A.J.*, **52,** 57 (1946).
HUANG, S.-S., *A.J.*, **61,** 49 (1956).
———, *J.R.A.S. Canada*, **51,** 91 (1957).
———, *Ap. J.*, **136,** 903 (1962*a*).
———, *A.J.*, **67,** 578 (1962*b*).
JEANS, J. H., *M.N.*, **85,** 2 (1924).
———, *ibid.*, p. 912 (1925).
JOY, A. H., *Pub. A.S.P.*, **54,** 35 (1942).
———, *ibid.*, **59,** 171 (1947).
KOPAL, Z., *Ap. J.*, **93,** 92 (1941).
———, *Close Binary Systems* (London: Chapman & Hall) (1959).
KUIPER, G. P., *Ap. J.*, **93,** 133 (1941).
MELTZER, A. S., *Ap. J.*, **125,** 359 (1957).
MOLTZ, L., *Ap. J.*, **115,** 562 (1952).
———, *ibid.*, **118,** 147 (1953).
SAHADE, J., *Stellar Atmospheres*, ed. J. L. Greenstein (Chicago: University of Chicago Press), ch. 12 (1960).
STRUVE, O., *Ap. J.*, **93,** 104 (1941).
———, *Stellar Evolution* (Princeton: Princeton University Press) (1950).
———, *Pub. A.S.P.*, **70,** 5 (1958).
SWINGS, P., *Zs. f. Ap.*, **12,** 40 (1936).
TISSERAND, F., *C.R.*, **120,** 125 (1895).
WOOD, F. B., *Ap. J.*, **112,** 196 (1950).
WOOLF, N. J., *M.N.*, **123,** 399 (1962*a*).
———, private communication (1962*b*).

RECENT ADVANCES CONCERNING THE EARLY HISTORY OF THE SOLAR SYSTEM

SU-SHU HUANG

Goddard Space Flight Center

The universe may be in a steady state as some cosmologists assert or it may be evolving as others are convinced; but there is very little doubt that chemical elements heavier than hydrogen were "cooked" inside stars. Therefore it may be reasonably assumed that in the very beginning our galaxy was formed primarily of hydrogen with little of heavier elements as they were yet to be formed in the stars. Consequently the first generation stars formed in our galaxy some 10 billion years or more ago were also made up originally of pure hydrogen.

As a result of energy loss through radiation at the surface, a star will contract if there is no other energy source inside the star besides the gravitational one. The contraction of the star releases the gravitational energy which supplies the energy that is lost at the surface and also raises the thermal energy content in the interior. Hence the internal temperature of the star steadily increases as it contracts until it becomes so hot (10^7 °K) that nuclear burning of hydrogen takes place in the central core.

The onset of thermonuclear reactions of converting hydrogen into helium inside the star arrests its gravitational contraction because now the energy released from nuclear reactions compensates completely the energy radiated away. This characterizes the main-sequence stage of stars. While this stationary state with a complete energy balance will last a very long time because of the prodigious reservoir of nuclear energy, the latter is nevertheless not limitless. Eventually the hydrogen in the central core will be exhausted, leaving only helium in it.

The helium does not enter into thermonuclear reactions at the temperature at which hydrogen is "burning." The reason for this lies in the different electric charges that hydrogen and helium nuclei carry. In order to find two atomic nuclei to enter a nuclear reaction, they must first be brought close together. Now all nuclei are positively charged and consequently repulse each other with a force proportional to their charges. To ensure nuclear reactions the physicist in the laboratory bombards one species of reacting nuclei with another of high speeds out of an accelerator. The high speed is to overcome the electrostatic force of repulsion so that the bombarding nuclei can penetrate the bombarded ones. In the stars no physicists are there to accelerate the nuclei. How could the nuclei be brought together to ensure reactions? It is thermal velocities that are responsible for overcoming the electrostatic repulsion. This is why the nuclear reactions in stars are called thermonuclear. Now the helium nuclei are doubly charged and exert stronger repulsive forces among themselves than do the hydrogen nuclei which are only singly charged. As a result higher velocities, and consequently in the case of stars, higher temperatures, are needed before helium nuclei enter into reaction than those required in the hydrogen "burning."

Therefore after the exhaustion of hydrogen in the central core, the nuclear flame will spread from the hydrogen-depleted core out to where hydrogen is still rich. Since the central core now has no nuclear energy source, it again contracts as the gravitational contraction is the natural tendency of a self-gravitating system. The gravitational energy released in the core will heat the helium core and raise its temperature to

where helium begins to enter thermonuclear reactions (10^8 °K). The net result of nuclear burning of "helium" is to convert helium to stable carbon. Thus, the elements that lie between helium and carbon in the periodic table are bypassed by helium "burning." But other species of atomic nuclei, it is now believed, can all be produced in successive stages of thermonuclear reactions that are taking place in the star as its temperature increases.

It first appears that formation of heavy elements should have nothing to do with the heavy elements that are now found in the solar system, especially in the terrestrial planets. Actually most elements heavier than hydrogen that constitute our planet and even our own bodies were manufactured in the stars in the manner we have just described. This is because at the late stage of stellar evolution when the temperatures reaches several billion degrees, the star will collapse. The physical reasons of stellar collapse at the late stage are too long to be described here. Suffice it to say that it is the result of a rapid loss in one form or another, of thermal energy of the star. A rapid loss of energy reduces the central pressure and upsets the pressure balance that is necessary to maintain a stable star. The result is an inevitable stellar collapse. Some astronomers very wittily compare the rapid energy loss in a star at this stage to a sudden wave of withdrawal of money from a bank, since both lead to catastrophe of no small degree.

When the central region of the star collapses, its outer part falls catastrophically inward as a cosmic avalanche, releasing a large amount of energy which causes the temperature of the infalling material to rise rapidly. The rise of temperature accelerates the thermonuclear reactions which in turn leads to a rapid release of nuclear energy. According to calculations such a cyclic process of temperature increase and energy release takes only a second or two before energy generated becomes so huge that a catastrophic explosion has to occur in the star. This kind of stellar explosion has indeed been observed in the supernova outbursts.

This violent event practically turns the star inside out and spreads all its heavy elements formed in its long history of evolution into space. In this way, the interstellar medium is continually enriched by chemical elements other than the primeval hydrogen. When the next generation of stars are formed from the interstellar clouds they will no longer be composed solely of hydrogen even at the very beginning of their existence. They will contain traces of heavy elements from the debris of the first-generation stars. Now except this difference in the original composition, the second generation stars will evolve along the path taken by the first generation stars. That is to say that more nuclear species heavier than proton will be formed by nuclear processes in the stars and spread in the interstellar media after explosion. In this way as more and more stars pass through the cycle of their life, more and more helium and heavier elements will be mixed in the interstellar clouds.

At this point it may be of interest to note that if we should like to use the word "evolution" in the same sense as geneticists use it in biological evolution, it is this change in the chemical compositions of the newly born stars from generation to generation that really characterizes evolution of stellar species, while stellar evolution in the sense we astrophysicists ordinarily mean should be called stellar growth and decay.

The enrichment of interstellar clouds by heavy elements through stellar explosion is not an idle speculation but has a sound support from observations. For it is now a well known fact of observation that there is a definite metal deficiency in the Population II stars which were formed in the early stage of our galaxy as compared with the Population I stars which have been formed only recently after eons of enrichment of heavy elements in the medium from which these new stars are formed.

Our solar system, formed some 4.5 billion years ago, must have been formed after many generations of stars exploded and spread their debris of heavy elements in the medium from which our solar system was emerged. Therefore there was a great deal of heavy elements in the day of its formation. Some of these heavy elements have found themselves in the earth and other terrestrial planets, in asteroids and meteorites and in comets and interplanetary grains. These of course are all old stories well known among astrophys-

icists who have an interest in the origin of the solar system. But recently a very interesting paper has been published by W. A. Fowler, J. Greenstein, and F. Hoyle of California Institute of Technology in Geophysical J. of RAS (6, 148–220, 1962) that throws new light on the early history of the solar system. It concerns the abundances of the nuclei of the isotopes D^2, Li^6, Li^7, Be^9, B^{10}, and B^{11}.

By the process of nucleosynthesis in stellar interiors as briefly described earlier in this article, it is very difficult to explain the presence of these nuclei. In the process of hydrogen "burning" in the main-sequence stars, D^2 is produced in equilibrium abundance in a ratio D^2/H^1 in the order of 10^{-17} which is many orders of magnitude smaller than the 1.5×10^{-4} found on the earth. Also, the isotopes of Li, Be, and B are bypassed in the nuclear synthesis in the stars by the formation of C^{12} from helium nuclei, as we have mentioned before. Even if they were formed by some unknown mechanism, they would still be rapidly destroyed in stars by nuclear reactions before and at the stage of hydrogen burning. Thus, Fowler and Burbidges suggested several years ago that these nuclear species are obtained in nuclear reactions of a nonthermonuclear character at the stellar surface. However the paper by Fowler, Greenstein, and Hoyle mentioned before gives a more definite scheme of production of these species and derives interesting results as regards the early history of the solar system from an investigation of their abundances.

Fowler, Greenstein, and Hoyle (hereafter Fowler et al.) assumed strong electromagnetic activities at the surface of the primeval sun. Particles, mainly protons, are accelerated by these activities to high energies. According to them, the formation of Li, Be, B can be obtained in the spallation process by these high-energy, nonthermal particles in interaction with relatively cool, moderate density material. Spallation is a kind of high-energy nuclear reaction that releases a large number of protons and neutrons. The protons and neutrons may be ejected, as the result of the reaction, individually or combined in the form of α-particles or even heavier nuclei. Thus, D, Li, Be, B and other nuclei may be ejected or remain as residues. In both cases they are said to be spallation products. That these nuclei can be formed in this way has been amply demonstrated by the production of them by spallation of heavy nuclei in cosmic radiation and in high-energy laboratories.

TABLE 1.—*Abundances and Cross Sections (Taken from W. A. Fowler, Science,* **135,** *1037, 1962)*

Isotope	Meteoritic or terrestrial abundance*	Spallation products (cross section in mbarn**)
H^1	4×10^3	
D^2	0.6	100
Li^6	7.4	15
Li^7	92.6	15
Be^9	20.0	11
B^{10}	4.5	38
B^{11}	19.5	11

*Based on the scale Si = 10^6.
**Millibarns = 10^{-27} cm^2.

For reasons that will be explained later the relevant spallation in the solar nebula occurred in an intermediate stage of the formation of terrestrial planets, namely after the planetary material had separated from the sun but before the planets themselves have been formed. It is assumed that solid condensations (planetesimals) were imbedded in a gaseous medium and were revolving around the sun. They were bombarded by high-energy particles, mainly protons, from the sun, producing these light elements as spallation products on their surface layers. The planetesimals will eventually coalesce to become the planets and the parent bodies of the meteorites.

Doubtless the high-energy particles generated on the surface of the young sun must have produced spallation products of Li, etc., in its own atmosphere. Thus, the fact discovered by Bonsack and Greenstein recently that Li is unusually abundant in the atmospheres of T Tauri stars gives a strong support to this theory of element formation, since T Tauri stars, just as the primitive sun, are stars still in the stage of gravitational contraction that leads to the main sequence.

Let us now consider clues to the early history of the solar system as derived by Fowler et al. table 1, which is taken from Fowler's paper in

Science (**135,** 1037, 1962), gives a comparison between the observed abundances and cross sections of spallation yields of the nuclear species concerned. The meteoritic or terrestrial abundances are listed in the second column according to the scale with $Si = 10^6$. Cross sections of spallation products are given in column 3 which should represent the relative yields of various nuclear species in spallation. When we compare the second and third column it can be easily seen that the two do not run in proportion as we would expect if spallation is the only reaction that involves these nuclei. The even-A isotopes Li^6 and B^{10} are very low in abundance relative to the odd-A isotopes Li^7, Be^9, and B^{11}. This, according to Fowler et al., is due to the fact that nuclei were subjected to slow neutron reactions simultaneously with the process of spallation. The even-A isotopes Li^6 and B^{10} have large cross sections to capture neutrons at low energy levels and thereby their abundances are lowered. On the other hand the odd-A isotopes Li^7, Be^9, and B^{11} do not react with neutrons at low energy; their abundances are not affected by thermal neutrons (i.e., neutrons at thermal velocities). This is a very convincing explanation for the discrepancy between the actual abundances and spallation cross sections, since neutrons are produced in the spallation processes in abundance.

From these considerations Fowler et al. are able to compute the neutron flux at the time when these light elements were formed as spallation products. The physical basis for this kind of calculation is easy to see. Let us take the case of B^{10} and B^{11}. As we have said before, the abundance of B^{11} is not affected by the presence of neutrons while that of B^{10} depends upon them. Therefore the ratio of the abundances of these two isotopes must be determined by the neutron flux times the time interval that B^{10} is exposed to neutron flux. In this way they derived a neutron flux of 1.3×10^7 neutrons per cm^2 per second if the time of spallation activity is taken to be 10^7 years. The flux will be higher if time is shorter than the assumed value because it is the product of the flux and time interval that is determined by calculation.

With this value of neutron flux, the Caltech scientists went on to compute the formation of deuterons as a result of capture of neutrons by protons. By the capture process as well as by direct production of deuterons in spallation they found an abundance ratio of D^2/H^1 in the irradiated material ten times the terrestrial value. This result points to the fact that only about 10% of the terrestrial material have been irradiated by thermal neutrons. Indeed this must be the case, for otherwise nuclear species with large cross sections for neutron capture such as Gd^{157} would have practically disappeared from the terrestrial material.

If only 10% of the terrestrial material has been irradiated by thermal neutrons, it is most likely that only the outer layers of the solid bodies formed by the material have been exposed to neutron flux.

Now, it can be found from the geometrical cross section for the nuclei in the planetesimals that the mean free path of high-energy protons is about 40 cm. Therefore if the bodies are spherical in shape we may take the outermost spherical shell of 40 cm in thickness as containing the material that has been irradiated by high-energy protons and subsequently by thermal neutrons. The condition that the mass in this shell is about 10% of the total mass of the spherical body gives directly the radius of the sphere to be about 10 meters.

On the other hand we may also argue that the radii of condensations at the time of spallation might vary over a wide range of values. Then the large bodies of asteroidal sizes such as suggested by R. A. Fish, G. G. Goles, and E. Anders of the University of Chicago would have suffered practically little effect percentagewise of spallation while bodies of radii less than 10 meters would have received its full share, making the over-all effect of irradiation at 10%. For this reason the value of about 10 meters as derived by Fowler et al. means only a sort of average value. What is sure from this study is that spallation took place after condensations had appeared in the solar nebula.

Fish, Goles and Anders further examined the energy for physical and chemical processing in the meteorites. They concluded that only extinct radioactivity seems to meet all requirements as the energy source and cited an example of Al^{26}

which has a mean life time of 10^6 years. Because of the short life time of those extinct radioactive elements it is evident that the parent bodies of the meteorites must have accreted within 10 million years after the nucleogenesis that produced these elements.

Other extinct radioactive nuclear species are I^{129} (J. H. Reynolds of Berkeley) and Pd^{107} (V. R. Murthy of La Jolla). They have mean life times of 2.5×10^7 and 10^7 years respectively. We have reasons to believe that both were present in some meteorites. Indeed Reynolds found that the Richardton stone meteorite is heavily enriched in Xe^{129} which is the decay product of I^{129}. All these examples show that the time from the nucleosynthesis to the formation of the parent bodies of meteorites could not have exceeded 10^6 to 10^8 years. In a longer time than these values these short-lived radioactive nuclides would have all decayed.

This short timescale for the formation raises a very serious difficulty as to how the parent bodies could have been formed so quickly. According to the theory of Fowler et al. the difficulty can be easily explained since it is assumed that I^{129}, Pd^{107} and Al^{26}, like Li, Be, etc. were directly formed in the parent bodies of terrestrial planets and meteorites.

It is doubtless that hydrogen was present during the synthesis of terrestrial deuterium, lithium, beryllium and boron. In the first place, the hydrogen is effective in bringing neutrons to thermal velocities by collisions. Also, the formation of deuterium requires the presence of hydrogen to capture neutrons. On the other hand hydrogen could not be as abundant as in its natural cosmic proportion. For if so, hydrogen would have captured all neutrons, leaving nothing for Li^6 and B^{10}. This would contradict the fact that Li^6 and B^{10} are much less abundant than Li^7 and B^{11} as shown in table 1. A quantitative examination by Fowler et al. indicates a hydrogen content of $H^1=8\times10^6$ on the $Si=10^6$ base. On the same scale, $H^1=3.2\times10^{10}$ in the sun and $H^1=4\times10^3$ on the earth at present. Therefore at the time when Li and other light elements were formed as spallation products, hydrogen was less abundant than it is in the sun but more abundant than it is on the earth.

From this consideration, Fowler et al. have concluded that the material that was to form the terrestrial planets had been irradiated by high-energy particles from the sun at the time when the material, in its course toward the formation of terrestrial planets, had nearly, but not yet completely separated from hydrogen. This shows that the terrestrial planets have not evolved by dissipation of hydrogen from very large protoplanets in which hydrogen was present originally in the same proportion as it was in the young sun. These conclusions agree with the results from other considerations (chemical and geophysical) that have been summarized in "The Planets" by Urey for example.

Fowler et al. further pointed out that $H=8\times10^6$ is close to twice the present value for terrestrial oxygen ($O=3.5\times10^6$). It could indicate that hydrogen and oxygen were trapped in the chemical combination of H_2O, presumably as ice, in the primitive planetesimals.

At about the same time nuclear clues were being sought for the formation of terrestrial planets, C. Hayashi of Kyoto University, during his stay at the Goddard Space Flight Center, studied the pre-main sequence evolution of stars in general and of the sun in particular. He has pointed out that the photospheric boundary condition must play an important role in pre-main sequence stars. As a result of this consideration he found that in the early phases of contraction, the star must be completely convective and its luminosity higher than that derived from radiative equilibrium, although the final approach to the main sequence still follows the models calculated some years ago by Henyey, LeLevier, and Levee. Because of the high luminosity the rate of gravitational contraction takes place more rapidly than we thought previously and the time scale of contraction is greatly shortened.

Not only does Hayashi's theory of pre-main sequence stars have a sound theoretical basis, it also explains several observational facts. One of them is connected with the abundance of nuclear species. It has long been known that Li is depleted in the sun while beryllium has a normal abundance. This indeed was a very puzzling fact. Now according to Hayashi's calculation the maximum temperature at the bottom of the convective

region in the contracting sun is just enough to destroy Li by nuclear reactions but somewhat too low to destroy Be.

Hayashi's classical works have induced many astrophysicists to turn their interest to the early stages of stars and the sun. One of the more recent works of such a nature is by J. Faulkner, K. Griffiths, and F. Hoyle of Cambridge University, which is published in the recent issue of M.N. (**126**, 1, 1963).

As we have seen earlier in this article, Fowler et al. require nuclear activities arising from thermal neutrons produced in the spallation process. Now Faulkner et al. observed that thermalization of neutrons seem to require the presence of hydrogen, not in gaseous form but in condensed form. Also, there is evidence that condensed water is indeed present in all meteorites. In fact Fowler et al. also suggested ice to be present in planetesimals. Now if all the activities from the spallation to the formation of the terrestrial planets were taking place in the contracting phase of the sun and if the sun at that stage was as luminous as is predicted by Hayashi, it would hardly be possible at the distance of one to three astronomical units from the sun for the condensation of water from the gaseous phase.

Faulkner et al. regard this discrepancy very serious and go on to examine Hayashi's result critically. They draw a list of factors that were not included in the work of Hayashi and R. Hōshi relating to contracting stars. Some of these factors concern the mathematical approximation, while others with improvement of opacity. When these factors are improved and calculations made, they have found that Hayashi's essential result still holds with only a small change in the evolutionary track of the star on the H–R diagram. Therefore Faulkner et al. consider the conflict between Hayashi's effect and the requirement of water condensation in the region now occupied by terrestrial planets unsolved and suggest that the presence of a strong magnetic field in the star may suppress the efficiency of convection and thereby reduce the luminosity. If so, the behavior of the early sun was inherently different from that of stars of approximately solar mass now observed in the young clusters. This suggestion may have some appeal for further investigation but is not too convincing as of now.

Actually, instead of the view taken by Faulkner et al. who challenge Hayashi's result, we may look at this conflict in a different light. The solar nebula may be opaque enough to ensure a low temperature at places where the terrestrial planets are now making their periodic revolution. Indeed, according to E. Opik of the University of Maryland and more recently to B. Donn of Goddard Space Flight Center, and G. W. Sears, of the Molecular Physics Laboratory at Rochester, N.Y., shielding of solar radiation by grains is very effective to lower the temperature in the solar nebula. Thus, while the sun might shine brightly in its early phase of evolution as Hayashi predicted, water could still be condensed from the gaseous phase to the liquid and even to solid phase in the regions of terrestrial planets. Therefore, contrary to Faulkner et al., we would conclude that the two investigations—the one by Fowler et al. on nuclear clues and the other by Hayashi on the solar history—have no basic conflict. In fact they represent two of the most significant advances in our understanding of the origin of the solar system in recent years.

TRANSFER OF ANGULAR MOMENTUM BETWEEN EJECTED PARTICLES AND THE BINARY SYSTEM

SU-SHU HUANG

Goddard Space Flight Center

Transfer of angular momentum between the ejected particle and the binary system itself has been studied numerically in the framework of the restricted three-body problem and the general behavior of such transfer for particles of high velocities of ejection is elucidated. Tables are given for the angular momentum that will be carried away to infinity under various initial conditions of ejection.

In the previous paper (Huang 1963) we have seen that the change in orbital period of a close binary system as a result of mass ejection by its components depends greatly on the angular momentum per unit mass, h_e, of the escaped particles. Therefore, it is important to know how much angular momentum is carried away per unit mass under any given initial condition. It is the purpose of this note to give, by numerical computation, some ideas about the behavior of angular-momentum transfer between the ejected particles and the binary system itself and to show that h_e of a particle at infinity, denoted by $h_{e,\infty}$, may be obtained for a given set of initial conditions by integrating the equations of motion of the particle over a relatively short period of time. All symbols used here have the same meanings as in the previous paper unless otherwise stated.

Since we are interested only in close binaries, it is permissible to set $e=0$. Then the equations of motion become identical with those in the restricted three-body problem. In this paper we shall investigate the transfer of angular momentum by the use of these equations.

Several authors, including Kuiper (1941), Kopal (1956, 1957), and Mrs. Gould (1957, 1959), have derived many orbits for gaseous particles in a close binary system from the solutions of the restricted three-body problem. One disadvantage of this kind of calculation is that it is impossible to compute the orbits and plot them for all conveivable initial conditions. Also, even if they were all plotted, it is difficult, if not impossible, to derive from these highly complicated and seemingly irregular orbits much information that may throw some light on the actual motion of gaseous particles in the system. Worst of all, such an approach necessarily neglects the collisions of particles themselves, although collisions cannot be neglected in this case (Pendergast, 1960). Consequently, with only a few exceptions, direct integration of the equations of motion has yielded few results of any great physical significance. However, as we shall see, useful information can be obtained by investigating the angular momentum of the particle as a function of time.

Let us choose a co-ordinate system (x, y) rotating with the circular motion of the binary stars around the center of mass of the system. Furthermore, we take the total mass of the system as the unit of mass, the separation between the two components as the unit of length, and $1/(2\pi)$ of the orbital period as the unit of time. Thus, if μ denotes the mass of one component that is located at $(1-\mu, 0)$, the other component will have mass $1-\mu$ and be located at $(-\mu, 0)$ in the rotating co-ordinate system.

The equations of motion can then be written in a dimensionless form (e.g., Moulton, 1914), which will not be given here. The angular mo-

mentum per unit mass in the present system of units, h_e, is now given by

$$h_e = x(dy/dt) - y(dx/dt) + x^2 + y^2. \quad (1)$$

From the equations of motion, we can easily show that

$$h_e = \int_{t_0}^{t} \frac{x_1 x_2 y}{r^3} \left\{ \left[1 - \left(\frac{2xx_2}{r^2} - \frac{x_2^2}{r^2} \right) \right]^{-3/2} - \left[1 - \left(\frac{2xx_1}{r_2} - \frac{x_1^2}{r^2} \right) \right]^{-3/2} \right\} dt, \quad (2)$$

where $r^2 = x^2 + y^2$ and $x_1 = -\mu$ and $x_2 = 1 - \mu$.

When the third body is far away from the system, the changes in x and in y are dominantly due to rotation of the co-ordinate system. Therefore, we may write as a first approximation,

$$x = r \cos t, \quad y = -r \sin t. \quad (3)$$

If we substitute x and y given by equations (3) in equation (2), we find that h_e assumes the following form:

$$\int_{t_0}^{t} \sin t \, dt f(r, \cos t).$$

If we regard r as constant, the integral vanishes when we integrate over a complete period of the binary motion, i.e., from $t = t_0$ to $t = t_0 + 2\pi$. Thus we derive the conclusion that no net angular momentum is transferred in a physically significant degree over a period when the particle is far away from the system.

If we expand the expressions in the square brackets in the integrand of the integral in equation (2) in terms of $1/r$, and take only the first term, we obtain

$$h_e = \frac{3\mu(1-\mu)}{4r^3} (\cos 2t_0 - \cos 2t), \quad (4)$$

if we again make use of equations (3) and set r as constant during integration. Thus the angular momentum of the particle at any moment follows a double sinusoidal curve of decreasing amplitude as r increases. The curve has maxima at $x = 0$ and minima at $y = 0$. Results of actual computations completely verify this prediction, which is based on our approximate calculation. In figure 1 we have illustrated for $t > 7.2$ the variation of angular momentum of the third body ejected from one of the components under the initial conditions

$$x = 0.85, \quad y = 0, \quad dx/dt = 4, \quad \text{and} \quad dy/dt = 0,$$

with $\mu = 0.3$. This kind of fluctuation in angular momentum is common to all particles that are escaping to infinity with reasonable speeds. Since the angular momentum which a particle under a given set of initial conditions will possess at infinity can be calculated, we now have only one value instead of an endless orbit to associate with a given set of initial conditions.

Another advantage of treating the angular momentum instead of the orbit itself comes from the consideration of collisions. In the case of the orbital method, a collision between two particles transforms the orbits to two completely different orbits. We can predict nothing about the behavior of the orbits before and after the collision. It is for this reason that the orbit approach faces its greatest weakness in treating the motion of gas in close binaries.

The situation becomes quite different when we use the angular-momentum approach because a definite relation exists that links the dynamical state of particles before and after collision. The total angular momentum of colliding particles is conserved in the process of collision. If many particles are involved in collision, the net result of interchange in angular momentum among the colliding particles tends to equalize the angular momenta of individual particles. Thus the chaotic motion of particles will turn more or less into streams. This is why gaseous rings can be formed by ejected matter. The angular momentum per unit mass in the stream can be obtained by taking the average of the same quantity for all the individual particles involved. In this way we may look upon the angular-momentum consideration as a link between the orbital approach and the hydrodynamic approach to the problem of stream motion in the binary system. While the present paper does not include a study of collision, it is evident that, for fast ejection, the results tabulated here can be applied, as an approximation, to aggregates of particles if we use the average values as the initial conditions.

We have computed the stabilized angular momentum for eight groups of sets of initial condi-

TABLE 1.—*Eight Groups of Initial Conditions*

No.	x	y	dx/dt	dy/dt	Type of ejection
1	$1-\mu$	r_2	0	V	Front.
2	$1-\mu$	$-r_2$	0	V	Rear.
3	$1-\mu+r_2$	0	V	0	Exterior.
4	$1-\mu-r_2$	0	V	0	Interior.
5	$-\mu$	$-r_1$	0	V	Front.
6	$-\mu$	r_1	0	V	Rear.
7	$-\mu-r_1$	0	V	0	Exterior.
8	$-\mu+r_1$	0	V	0	Interior.

tions that are given in Table 1, where r_1 and r_2 represent, respectively, the radii of the two components. These groups include the cases of ejection in four mutual perpendicular directions for each component, as the integration has been confined in the $x-y$ plane. The computation was carried out on the 7090 IBM computer at the Goddard Space Flight Center, the machine program being written and actual computation supervised by Clarence Wade, Jr. An accuracy of at least six significant figures was maintained throughout, as judged by the constancy of the Jacobian constant.

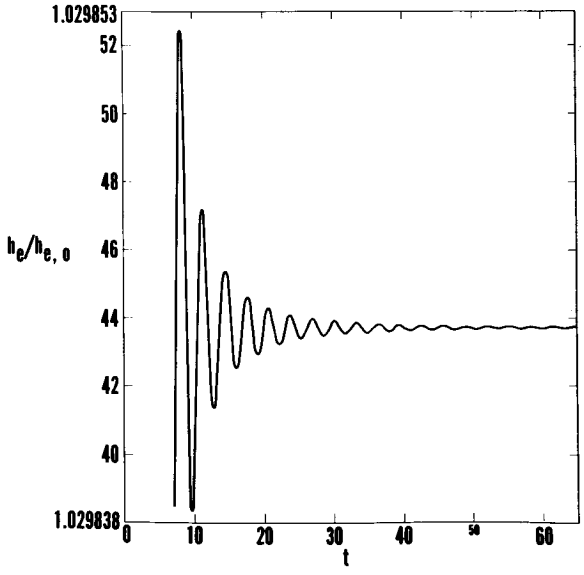


FIGURE 1.—The variation in angular momentum of a particle escaping from a binary system. It is the typical behavior of an escaping particle that its angular momentum undergoes damped oscillations before reaching a value that it will carry to infinity.

In most cases an accuracy of eight significant figures has been achieved over the entire range of integration.

As may be expected intuitively, the transfer of angular momentum from the binary system to the particle takes place most appreciably when the particle is near to one of the components, and this is verified by our extensive computations. Therefore, one cannot easily predict the angular momentum of a particle as a function of the time when it is confined to the system.

Our interest, however, concerns only the escaped particles, for which the pattern of transfer is relatively simple. In the case of ejection from the front of the star (i.e., in the direction of its orbital motion), the angular momentum of the ejected particle starts from a positive value, decreases gradually, and then stabilizes to a smaller positive value after a series of fluctuations mentioned before and shown in figure 1. In the case of ejection from the rear side of the star (i.e., against the direction of its orbital motion), its angular momentum is negative at the time of ejection. It increases through the interaction with the binary system itself and, as in the other case, stabilizes through fluctuations to a negative value. Thus, in both cases, the continuous interaction after ejection neutralizes a part of the angular momentum acquired at the time of ejection.

In the case of outward ejection along the x-axis (positive for the component with mass μ and negative for the other component), the angular momentum initially has a positive value, increases gradually, and then stabilizes, after damped oscillations, to a constant value without further significant change afterward. Thus a par-

ticle finally possesses more angular momentum than its initial value. In all cases the faster the initial velocity of ejection, the shorter the time for the particle to reach the stabilized value of angular momentum. Also, the angular momentum is stabilized to a value nearer to its initial value in the case of a fast ejection than in the case of slower ejections.

TABLE 2.—*Angular Momentum of an Escaped Particle as a Function of Initial Conditions*

No. of group	V		$\mu=0.1$	$\mu=0.2$	$\mu=0.3$	$\mu=0.4$	$\mu=0.5$
1 ($r_2=0.15$)	4	$h_{e,0}$	4.4325	3.8625	3.3125	2.7825	2.2725
		$h_{e,\infty}$	4.299	3.620	2.987	2.402	1.8669
	6	$h_{e,0}$	6.2325	5.4625	4.7125	3.9825	3.2725
		$h_{e,\infty}$	6.145	5.305	4.503	3.742	3.019
	10	$h_{e,0}$	9.8325	8.6625	7.5125	6.3825	5.2725
		$h_{e,\infty}$	9.780	8.569	7.390	6.242	5.125
	15	$h_{e,0}$	14.3325	12.6625	11.0125	9.3825	7.7725
		$h_{e,\infty}$	14.298	12.600	10.931	9.290	7.676
2 ($r_2=0.15$)	−4	$h_{e,0}$	−2.7675	−2.5375	−2.2875	−2.0175	−1.7275
		$h_{e,\infty}$	−2.647	−2.320	−1.9970	−1.6802	−1.3703
	−6	$h_{e,0}$	−4.5675	−4.1375	−3.6875	−3.2175	−2.7275
		$h_{e,\infty}$	−4.486	−3.991	−3.494	−2.995	−2.493
	−10	$h_{e,0}$	−8.1675	−7.3375	−6.4875	−5.6175	−4.7275
		$h_{e,\infty}$	−8.118	−7.249	−6.371	−5.484	−4.588
	−15	$h_{e,0}$	−12.6675	−11.3375	−9.9875	−8.6175	−7.2275
		$h_{e,\infty}$	−12.635	−11.277	−9.910	−8.528	−7.134
3 ($r_2=0.15$)	4	$h_{e,0}$	1.1025	0.9025	0.7225	0.5625	0.4225
		$h_{e,\infty}$	1.1107	0.9178	0.7440	0.5890	0.4523
	6	$h_{e,0}$	1.1025	0.9025	0.7225	0.5625	0.4225
		$h_{e,\infty}$	1.1065	0.9097	0.7322	0.5739	0.4348
	10	$h_{e,0}$	1.1025	0.9025	0.7225	0.5625	0.4225
		$h_{e,\infty}$	1.1040	0.9053	0.7262	0.5667	0.4270
	15	$h_{e,0}$	1.1025	0.9025	0.7225	0.5625	0.4225
		$h_{e,\infty}$	1.1033	0.9038	0.7242	0.5645	0.4525
4 ($r_2=0.15$)	−4	$h_{e,0}$	0.5625	0.4225	0.3025	0.2025	0.1225
		$h_{e,\infty}$	0.8265	0.8954	0.9264	0.9172	0.8659
	−6	$h_{e,0}$	0.5625	0.4225	0.3025	0.2025	0.1225
		$h_{e,\infty}$	0.8201	0.8819	0.9067	0.8936	0.8421
	−10	$h_{e,0}$	0.5625	0.4425	0.3025	0.2025	0.1225
		$h_{e,\infty}$	0.8156	0.8730	0.8941	0.8788	0.8269
	−15	$h_{e,0}$	0.5625	0.4225	0.3025	0.2025	0.1225
		$h_{e,\infty}$	0.8138	0.8693	0.8891	0.8730	0.8209

TABLE 2.—Continued

No. of group	V		$\mu=0.1$	$\mu=0.2$	$\mu=0.3$	$\mu=0.4$	$\mu=0.5$
5 ($r_1=0.2$)	−4	$h_{e,0}$	0.4500	0.8800	1.3300	1.8000	2.2900
		$h_{e,\infty}$	0.3414	0.6907	1.0862	1.5262	2.010
	−6	$h_{e,0}$	0.6500	1.2800	1.9300	2.6000	3.2900
		$h_{e,\infty}$	0.5839	1.1633	1.7779	2.427	3.111
	−10	$h_{e,0}$	1.0500	2.0800	3.1300	4.2000	5.2900
		$h_{e,\infty}$	1.0121	2.013	3.042	4.010	5.186
	−15	$h_{e,0}$	1.5500	3.0800	4.6300	6.2000	7.7900
		$h_{e,\infty}$	1.5252	3.036	4.572	6.134	7.721
6 ($r_1=0.2$)	4	$h_{e,0}$	−0.3500	−0.7200	−1.0700	−1.4000	−1.7100
		$h_{e,\infty}$	−0.2598	−0.5613	−0.8639	−1.1669	−1.4697
	6	$h_{e,0}$	−0.5500	−1.1200	−1.6700	−2.2000	−2.7100
		$h_{e,\infty}$	−0.4909	−1.0154	−1.5334	−2.045	−2.549
	10	$h_{e,0}$	−0.9500	−1.9200	−2.8700	−3.8000	−4.7100
		$h_{e,\infty}$	−0.9147	−1.8572	−2.788	−3.706	−4.613
	15	$h_{e,0}$	−1.4500	−2.9200	−4.3700	−5.8000	−7.2100
		$h_{e,\infty}$	−1.4264	−2.878	−4.315	−5.737	−7.144
7 ($r_1=0.2$)	−4	$h_{e,0}$	0.0900	0.1600	0.2500	0.3600	0.4900
		$h_{e,\infty}$	0.1004	0.1773	0.2713	0.3830	0.5126
	−6	$h_{e,0}$	0.0900	0.1600	0.2500	0.3600	0.4900
		$h_{e,\infty}$	0.0940	0.1669	0.2589	0.3699	0.5000
	−10	$h_{e,0}$	0.0900	0.1600	0.2500	0.3600	0.4900
		$h_{e,\infty}$	0.0914	0.1625	0.2532	0.3637	0.4938
	−15	$h_{e,0}$	0.0900	0.1600	0.2500	0.3600	0.4900
		$h_{e,\infty}$	0.0906	0.1611	0.2515	0.3617	0.4918
8 ($r_1=0.2$)	4	$h_{e,0}$	0.0100	0.0000	0.0100	0.0400	0.0900
		$h_{e,\infty}$	0.2968	0.5220	0.7059	0.8428	0.9296
	6	$h_{e,0}$	0.0100	0.0000	0.0100	0.0400	0.0900
		$h_{e,\infty}$	0.2976	0.5148	0.6891	0.8184	0.9020
	10	$h_{e,0}$	0.0100	0.0000	0.0100	0.0400	0.0900
		$h_{e,\infty}$	0.2948	0.5073	0.6768	0.8027	0.8848
	15	$h_{e,0}$	0.0100	0.0000	0.0100	0.0400	0.0900
		$h_{e,\infty}$	0.2933	0.5040	0.6720	0.7968	0.8783

The case of inner ejection along the x-axis (i.e., negative velocity for the μ component and positive velocity for the $1-\mu$ component) is more complicated than the other cases, as we would expect. Actual computations show that it differs from the other cases mainly when the velocity of ejection is high, for then the particle will either simply fall into the companion component or be

strongly perturbed by it during a close encounter such that the angular momentum is greatly modified. Thus we cannot make a general statement about the angular momentum for high velocities of ejection. However, it is interesting to note that for intermediate velocities, say between 4 and 15 in our units, the stabilized value of angular momentum does not vary greatly with the initial velocity.

Some of the numerical results obtained for various sets of initial conditions are given in table 2 according to the order listed in table 1. For the first four groups we have adopted $r_2 = 0.15$, while for the last four groups $r_1 = 0.20$. Five values for μ and four values for V have been used for each group. There are two entries for each combination of μ and V, the first one being the initial angular momentum $h_{e,0}$ corresponding to the initial conditions and the second being the stabilized value of angular momentum, i.e., $h_{e,\infty}$. The values $h_{e,\infty}$ in most cases were obtained after we integrated the equations of motion up to $t = 20$. But in some cases where the convergence is slow, we have to reach $t = 50$ before a stabilized value can be determined accurately.

Only velocities greater than 4 are included in the table because it is difficult to present a clear-cut picture of the ejected particle even when the result is represented in terms of angular momentum. In other words, the present analysis does not help much the problem of gaseous motion inside the close binary system, but, whem combined with the results obtained in the previous paper, it does help us understand the effect of ejection of mass to infinity on the orbital period of the binary. For higher velocities of ejection than the listed values, the particle will practically carry the initial value of angular momentum to infinity, except in the case of interior ejection.

It is a pleasure to express my sincere thanks to Mr. Clarence Wade, Jr., who performed the integration on the IBM 7090 computer at our Center and to Mrs. Priscilla Week, who plotted, in the course of the present investigation, a number of curves, including the one presented here.

REFERENCES

GOULD, N. L., *Pub. A.S.P.*, **69,** 541 (1957).
———, *A.J.*, **64,** 136 (1959).
HUANG, S.-S., *Ap. J.*, **138,** 471 (1963).
KOPAL, Z., *Ann. d'ap.*, **19,** 298 (1956).
———, *Non-stable Stars*, ed. G. H. Herbig (Cambridge: Cambridge University Press), chap. 17 (1957).
KUIPER, G. P., *Ap. J.*, **93,** 133 (1941).
MOULTON, F. R., *An Introduction to Celestial Mechanics* (2d ed.; New York: Macmillan Co.), chap. 8 (1914).
PRENDERGAST, K. H., *Ap. J.*, **132,** 162 (1960).

PRELIMINARY STUDY OF PERIODIC ORBITS OF INTEREST FOR MOON PROBES. II

SU-SHU HUANG AND CLARENCE WADE, JR.

Goddard Space Flight Center

Under the approximation of the restricted three-body problem, two families of periodic orbits that enclose both the earth and the moon in the plane of the earth-moon orbit have been derived and their initial conditions tabulated. Their stability is examined by investigating the variation of the difference of two successive periods with cycles. In the last section, two sequences of initial conditions have been studied in order to find some periodic orbits in the three-dimensional case. Among those investigated, none with the desired characteristic, namely of enclosing both the earth and the moon, has been found.

TWO FAMILIES OF DESIRED ORBITS

In a previous paper (Huang, 1962) in which some of the symbols used here have been defined, we have studied, under the approximation of the restricted three-body problem and by means of numerical processes, some interesting orbits which may provide a useful background for deriving periodic orbits for the moon-probing vehicle in the actual earth-moon-sun system. While it has been pointed out that two families of periodic orbits exist—one of direct motion and the other of retrograde motion—for the case $P/P_0=\frac{2}{3}$, only one orbit in each family has actually been given in the previous paper. In the present section we shall give a series of orbits in each family, which are obtained with the aid of an IBM 7090 computer.

Previously, the method of successive approximation was based on the idea that the third body, in a periodic orbit, should recover its initial position and velocity after a certain period of time. Such an approach, though direct and obvious, is not the most efficient. A new scheme is thereby introduced by which the desired orbits can be derived rapidly.

In order to see this new scheme, let us define the period p of a *nearly* periodic orbit by the time interval between two successive crossings over the x axis by the third body at points near to each other. It is apparent that the period thus defined for a nearly periodic orbit changes from cycle to cycle. Thus, we may denote p_1, p_2, p_3, $\cdots$, p_n, $\cdots$ as the periods of different cycles. All of them may be obtained by interpolation from the results of integration of the equations of motion. Needless to say, we should have $p_1=p_2=p_3=\cdots=p_n=\cdots$ for the true periodic orbits. Thus the following quantity:

$$\Delta_{n+1,n}=p_{n+1}-p_n, \tag{1}$$

measures the deviation from the periodic orbit.

Following the previous paper, we specify the initial conditions as

$$x=x_0, \quad y=0, \quad \dot{x}=0, \quad \dot{y}=\dot{y}_0, \tag{2}$$

and integrate the equations of motion many times for a fixed value of x_0 but with a series of values for $\dot{y}_0$ which are different from one another only slightly. For each $\dot{y}_0$ we obtain a value $\Delta_{2,1}$ after integrating the equations up to $t>p_1+p_2$. The desired orbit is obtained by finding the value of $\dot{y}_0$ which makes $\Delta_{2,1}$ vanish, a process being performed by interpolation. Since $\Delta_{2,1}$ is very sensitive to the change in $\dot{y}_0$, the process of successive approximation operated in this way converges rapidly. Thus, we have determined six periodic orbits for each family. They are so chosen that when the third body is on the far side of the moon,

TABLE I.—*Initial Conditions for Periodic Orbits in Direct Motion*

x_0	$\dot{y}_0$	p_1	$\Delta_{2,1}$
−0.39215	−1.6102480	11.643597	−0.000457
	−1.6102481	11.643566	−0.000380
−0.37215	−1.7014694	11.780075	−0.000091
	−1.7014695	11.780050	+0.000453
−0.35215	−1.7982280	11.900590	−0.000175
	−1.7982281	11.900571	+0.000161
−0.33215	−1.9013902	12.004325	−0.000025
	−1.9013903	12.004311	+0.000178
−0.31215	−2.0120181	12.092292	−0.000105
	−2.0120182	12.092282	−0.000014
−0.29215	−2.1314275	12.166330	−0.000034
	−2.1314276	12.166323	+0.000036

the closest distance between them lies between 0.08 and 0.20. The results of computation are given in table I for the direct orbits and in table II for the retrograde orbits. The integration as performed by the fourth-order Runge–Kutta method with double precision on the IBM 7090, with $\Delta t = 0.0078125$. Therefore, the constant of motion maintains its constancy in every case for at least six significant figures for a time interval of 4π. Lagrange four-point inverse interpolation was then used to compute the value of p_n from the integrated values of $y(t)$.

TABLE II.—*Initial Conditions for Periodic Orbits in Retrograde Motion*

x_0	$\dot{y}_0$	p_1	$\Delta_{2,1}$
−0.49215	2.1636769	12.836882	+0.000002
	2.1636770	12.836887	−0.000013
−0.47215	2.1937505	12.791218	+0.000009
	2.1937506	12.791223	−0.000005
−0.45215	2.2272677	12.755097	+0.000004
	2.2272678	12.755102	−0.000011
−0.43215	2.2645059	12.726637	+0.000015
	2.2645060	12.726642	0.000000
	2.2645061	12.726648	−0.000017
−0.41215	2.3058168	12.704262	+0.000010
	2.3058169	12.704268	−0.000007
−0.39215	2.3516410	12.686682	+0.000009
	2.3516411	12.686689	−0.000011

For each value of x_0 in both tables, two values of $\dot{y}_0$ are given—one with $\Delta_{2,1}$ being positive and the other with $\Delta_{2,1}$ negative. Therefore, the correct value of $\dot{y}_0$ for each x_0 that corresponds to the periodic orbit must lie somewhere between these two tabulated values. Here we can clearly see how sensitive $\Delta_{2,1}$ is with respect to the change of $\dot{y}_0$.

TABLE III.—*Variation in Period for the Direct Orbit* ($x_0 = -0.31215$).

n	$\dot{y}_0 = -2.0120181$		$\dot{y}_0 = -2.0120182$	
	$\Delta_{n+1,n}$	$\left\lvert\frac{\Delta_{n+1,n}}{\Delta_{n,n-1}}\right\rvert$	$\Delta_{n+1,n}$	$\left\lvert\frac{\Delta_{n+1,n}}{\Delta_{n,n-1}}\right\rvert$
1	−0.000105		+0.000014	
2	+0.000909	8.7	−0.000131	9.4
3	−0.007777	8.6	+0.001124	8.6
4	+0.066079	8.5	−0.009618	8.6
5	−0.644121	9.7	+0.081658	8.5
6			−0.841578	10.3

Needless to say, the present procedure does not provide an analytic proof of the existence of the computed families of orbits. However, it does give an intuitive assurance that such families may perhaps exist in a rigorous sense.

STABILITY CONSIDERATIONS

The stability of a periodic orbit in the restricted three-body problem may be learned from what Poincaré has called characteristic exponents in the solution of the Hill equation which has been recently studied by Message (1959). It is now easy to study it. Here we shall present a more intuitive and much easier way for demonstrating the stability or the instability of a periodic orbit by investigating the change in period, i.e., $\Delta_{n+1,n}$, of the $(n+1)$th cycle from the nth cycle.

If a periodic orbit is stable, any orbit in its neighborhood deviates from it always by a small amount as time increases. Consequently, $\Delta_{n+1,n}$ must oscillate without increasing its amplitude with respect to the increase of n (which is equivalent to the increase in t). On the other hand, if the amplitude of $\Delta_{n+1,n}$ increases with n, the periodic orbit cannot be stable. Hence the problem of stability reduces to one of investigating the behavior of $\Delta_{n+1,n}$ with respect to n, which can be, of course, obtained in the course of numerical integration of the equations of motion.

In the previous paper we have demonstrated by graphs that the direct orbit is unstable while the retrograde orbit is stable. Here we shall illustrate quantitatively the instability of the direct orbit and the stability of the retrograde orbit by tabulating $\Delta_{n+1,n}$. Table III lists the successive periods p_n for two values of $\dot{y}_0$ in the next to the last case in Table I. Here we see $\Delta_{n+1,n}$ fluctuates from positive to negative and vice versa with increasing amplitudes until its magnitude is so large that the orbit can no longer be regarded as nearly periodic. Indeed, the increase in amplitude of $\Delta_{n+1,n}$ with respect to n can be fairly represented by an exponential function as the ratio of the two consecutive Δ's, i. ., $|\Delta_{n+1,n}/\Delta_{n,n-1}|$ as given in the third and fifth column is approximately equal and is greater than one. This clearly shows the characteristic of instability.

On the other hand, the periodic orbits in retrograde motion are stable, because $\Delta_{n+1,n}$ oscillates with a nearly constant amplitude. For example, the variation in $\Delta_{n+1,n}$ with n for the case ($x_0 = -0.43215$) which corresponds to the fourth case given in table II, is listed in table IV for four different values of $\Delta\dot{y}_0$ and where $\Delta\dot{y}_0$ represents the deviation of $\dot{y}_0$ from the correct value of the periodic orbit. For each $\Delta\dot{y}_0$, the oscillation of $\Delta_{n+1,n}$ does not increase in amplitude as n or t increases. Moreover, the amplitude becomes smaller and smaller as the deviation $\Delta\dot{y}_0$ decreases. This is, of course, the characteristic of a stable orbit.

THREE-DIMENSIONAL CASE

We have tried to find some periodic orbits on which the third body encounters the moon at regular intervals. The introduction of the third dimension into the problem broadens the choice of the initial conditions and consequently complicates the processes of finding the desired orbits. We have, therefore, to limit our choice of initial conditions to the following combinations:

$$x = x_0, \quad y = 0, \quad z = 0,$$
$$\dot{x} = 0, \quad \dot{y} = \dot{y}_0, \quad z = z_0. \tag{3}$$

TABLE IV.—*Variation in period* ($\Delta_{n+1,n}$) *for the retrograde orbit* ($x_0 = -0.43215$)

$\dot{y}_0$	2.2645060	2.2645065	2.2645070	2.2645075
np_1	12.726642	12.726670	12.726698	12.726726
1	0.000000	−0.000079	−0.000158	−0.000237
2	+0.000002	+0.000082	+0.000164	+0.000246
3	−0.000003	−0.000006	−0.000012	−0.000018
4	−0.000001	−0.000077	−0.000152	−0.000227
5	+0.000004	−0.000087	+0.000169	+0.000252
6	−0.000003	−0.000013	−0.000021	−0.000034
7	+0.000002	−0.000073	−0.000149	−0.000217
8	0.000000	+0.000089	+0.000177	+0.000259
9	−0.000003	−0.000022	−0.000039	−0.000054
10	−0.000001	−0.000066	−0.000136	−0.000202
11	+0.000006	+0.000089	+0.000178	+0.000263
12	−0.000005	−0.000027	−0.000046	−0.000072
13	+0.000001	−0.000059	−0.000130	−0.000186

TABLE V.—*First Sequence of Initial Conditions which Lead to* $p_1(z=0)=p_1(y=0)(x_0=-0.39215)$

$\dot{y}$	$\dot{z}$	$x(y=z=0)$	p_1
0.075	1.88403	−0.3916	12.5757
0.105	1.75045	−0.3921	12.5655
0.135	1.75491	−0.3921	12.5661
0.165	1.82376	−0.3885	12.5583
0.195	1.89940	−0.3917	12.5772
0.225	1.97678	−0.4169	12.5408
0.255	1.97886	−0.4167	12.5419
0.285	1.98049	−0.4165	12.5428
0.315	1.98166	−0.4164	12.5438
0.345	1.98237	−0.4162	12.5448
0.375	1.98263	−0.4160	12.5457
0.405	1.98244	−0.4158	12.5467
0.435	1.98179	−0.4156	12.5476
0.465	1.90610	−0.3919	12.5810
0.495	1.97913	−0.4151	12.5494
0.525	1.83067	−0.3891	12.5653
0.555	1.69759	−0.3920	12.5688
0.585	1.63092	−0.3922	12.5694
0.615	1.75632	−0.3921	12.5756
0.645	1.75205	−0.3921	12.5762
0.675	1.88475	−0.3921	12.5846

Even with this limitation we still find that there are too many initial conditions to be studied. We have chosen among them two sequences given by

$$(1)\quad x_0=-0.39215,$$

$$(2)\quad x_0=-\dot{y}_0. \tag{4}$$

In the first case the sequence is obtained by varying $\dot{y}_0$ while in the second case by varying x_0. The second sequence corresponds to the launching of the third body with a velocity perpendicular to the plane of orbit of the moon.

TABLE VI.—*Second Sequence of Initial Conditions which lead to* $p_1(z=0)=p_1(y=0)$

y_0	$\dot{z}_0$	$x(y=z=0)$	p_1
−0.39215	1.98258	−0.4159	12.5463
−0.37215	2.05440	−0.3927	12.5479
−0.35215	2.13174	−0.3695	12.5495
−0.33215	2.21547	−0.3467	12.5510
−0.31215	2.24207	−0.3120	12.5850
−0.29215	2.34458	−0.2921	12.5863

Generalizing the procedure we have proposed in the first section, we now define a period of the nth cycle with respect to the xy plane, denoted by $p_n(z=0)$, as the time interval between the $(n+1)$th and the nth passage of the third body through the xy plane at x near to the initial value x_0. Similarly, we define the period of the nth cycle with respect to the xz plane, denoted by $p_n(y=0)$ as the interval between the $(n+1)$th and the nth passages of the third body through the xz plane near to x_0.

Let us now consider the first sequence by assigning numerical values to $\dot{y}_0$ and $\dot{z}_0$, and integrate the equations of motion. It is evident that for an arbitrary pair of values for $(\dot{y}_0, \dot{z}_0)$, $p_1(z=0)\neq p_1(y=0)$. Therefore, we fix $\dot{y}_0$ and vary $\dot{z}_0$ until

$$p_1(z=0)=p_1(y=0) \tag{5}$$

is satisfied. In this way one value of $\dot{z}_0$ is obtained for each fixed value of $\dot{y}_0$. Similarly, we have found one value z_0 for each given value of x_0 in the second sequence that leads to Eq. (5). In tables V and VI we have given the results of our computation. All pairs of initial conditions given here yield to orbits that satisfy Eq. (5). Since this is only a probing investigation of periodic orbits for moon probes in the three-dimensional case, computation was performed by the machine only with single precision. Consequently fewer significant figures are given in these two tables than in other tables in this paper.

The condition given by Eq. (5) is not sufficient to warrant a periodic orbit. In fact, most are not. In order to be periodic they must furthermore satisfy either the conditions $p_1(z=0)=p_1(y=0)=4\pi$ and $x(t=4\pi)=x_0$ or the conditions

$$p_1(z=0)=p_1(y=0)=p_2(z=0)=\cdots$$
$$=p_n(z=0)=p_n(y=0)=\cdots.$$

We have examined only cases which are to satisfy the first set of conditions, i.e., the period is equal to exactly two complete revolutions of the moon around the earth. These conditions make the actual search for periodic orbits in the three-dimensional case extremely tedious. On the other hand, if a periodic orbit should be derived in the framework of the restricted three-body problem, perhaps we could expect that a similar orbit may

be derived in the actual system of the earth and the moon by modifying the initial conditions.

When we examine the variations of p_1 we find that some initial conditions must exist which lead to

$$p_1(z=0)=p_1(y=0)=4\pi.$$

For example, the desired initial conditions perhaps lie near $\dot{y}_0=0.135$ in the first sequence. However, the orbits which are obtained by these initial conditions do not pass around the far side of the moon, although it appears that they are stable in the sense defined in the previous paper.

For the second sequence we obtain two values of $\dot{z}_0$ for each x_0 that lead to $p_1(y=0)=p_1(z=0)$ but neither of them is equal to 4π for the range of x (from -0.29215 to -0.39215) considered.

REFERENCES

HUANG, S.-S., *Astron. J.*, **67**, 304, 1962.
MESSAGE, P. J., *ibid.*, **64**, 226, 1959.

MODELS FOR STARS OF VERY LOW MASS

SHIV S. KUMAR*

Goddard Space Flight Center

Completely convective models have been constructed for stars of masses 0.09, 0.08, 0.07, 0.06, 0.05, and 0.04 (solar units), taking into account the nonrelativistic degeneracy of the stellar material. The properties of these models are presented in tabular form and in the form of graphs. It is shown that there is a lower limit to the mass of a main sequence star. The stars with mass less than this limit became completely degenerate stars or "black" dwarfs as a consequence of the gravitational contraction and therefore they never go through the normal stellar evolution.

INTRODUCTION

Very little observational or theoretical information is available concerning the structure of stars of very low mass. In this paper an attempt will be made to study the internal structure of stars of mass $M < 0.1 M\odot$ by using suitable theoretical models. In particular we shall be concerned with the effects of degeneracy on the structure of stars having masses between $0.09 M\odot$ and $0.04 M\odot$.

THE EQUATION OF STATE FOR THE STELLAR MATERIAL

We shall use here the following equation os state for the partially degenerate matter (Tolman, 1938):

$$P = P_n + P_e = N_n kT + N_e \frac{V(\alpha, 3/2)}{V(\alpha, 1/2)} kT , \quad (1)$$

where P is the total gas pressure, P_n is the pressure due to nuclei which are assumed to be nondegenerate, P_e the pressure due to electrons, N_n the density of nuclei, N_e the density of the electrons, k the Boltzmann constant, T the temperature, and the function $V(\alpha, \rho)$ is defined by

$$V(\alpha, \rho) = \frac{1}{\Gamma(\rho + 1)} \int_0^\infty \frac{z^\rho \, dz}{e^{\alpha + z} + 1}, \quad (2)$$

where α is a parameter which is a function of N_e and the temperature, and ρ is equal to $\frac{3}{2}$ or $\frac{1}{2}$.

*Dr. Kumar was a National Academy of Sciences-National Research Council Postdoctoral Research Associate with the NASA. He is now with the Physical Research Laboratory, Ahmedabad-9, India.

The equation of state (equation 1) is derived by using the Fermi-Dirac statistics and is valid for a mixture of nuclei and electrons in which the velocities of electrons are small as compared with the velocity of light. At high densities, the electron gas becomes degenerate but the degeneracy will be nonrelativistic in the stars of low mass because we do not expect to have densities higher than 10^4 gm/cm^3 in these objects.

MODELS FOR STARS OF VERY LOW MASS

In order to evaluate physical quantities such as the central temperature and the central density, we have to make use of certain models for the stars under study. Now, we are primarily concerned with these stars when they are contracting or when nuclear reactions involving the destruction of H^2, Li^6, Li^7, and B_e^9 are taking place. In both cases we can assume that the models are completely convective, as has recently been shown by Hayashi (1962). Therefore, we study the structure of stars of low mass by assuming that they can be represented by spheres of polytropic index 1.5. Before we derive the expressions for the temperature, density, and pressure inside the star, let us introduce here, following Tolman's notation, the degeneracy parameter y given by

$$y = \frac{N_e h^3}{2(2\pi m_e kT)^{3/2}} . \quad (3)$$

It can be shown that

$$y = V(\alpha, 1/2). \quad (4)$$

Denoting $V(\alpha, \frac{3}{2})/V(\alpha, \frac{1}{2})$ by $D(y)$, we obtain from equation 1:

$$T = \frac{\mu M_H}{k} \frac{P}{\rho} \frac{N_n + N_e}{N_n + D(y) N_e} , \quad (5)$$

where ρ is the density and μ is the molecular weight defined by

$$\frac{1}{\mu} = 2X + \frac{3}{4} Y + \frac{1}{2} Z . \quad (6)$$

Here X, Y, and Z are the abundances, by weight, of hydrogen, helium, and heavier elements, respectively. If we write

$$\Lambda(y) = \frac{N_n + N_e}{N_n + D(y) N_e} , \quad (7)$$

then the equation of state (equation 5) takes the form

$$T = \frac{\mu M_H}{k} \frac{P}{\rho} \Lambda(y) . \quad (8)$$

This form of the equation is similar to the equation of state for the perfect gas:

$$T = \frac{\mu M_H}{k} \frac{P}{\rho} . \quad (9)$$

Because of degeneracy of the electron gas, we have an additional factor $\Lambda(y)$ in equation 8. Using the equation of state (equation 8) and the fact that the star is in convective equilibrium we obtain the following expressions for the temperature, density, and pressure at any point inside the star:

$$T = 1.246 \times 10^7 \mu \frac{M}{R} \Lambda(y) \theta(\xi) , \quad (10)$$

$$\rho = 8.446 \frac{M}{R^3} \theta(\xi)^{3/2} , \quad (11)$$

$$P = 8.680 \times 10^{15} \frac{M^2}{R^4} \theta(\xi)^{5/2} , \quad (12)$$

where $\theta(\xi)$ is the Lane–Emden function for the polytrope of index 1.5 and

$$\xi = \frac{r}{1.905 \times 10^{10} R} , \quad (13)$$

r being the distance in centimeters measured from the center. Here M and R are expressed in solar units. Because the models chosen for study are in convective equilibrium, the quantity $\Lambda(y)$ is a constant throughout the star. The fact that y or $\Lambda(y)$ has the same value in all regions of the star is something unique to the completely convective models, as has already been pointed out by Limber (1958) who studied completely convective models for M dwarfs. It is this fact that greatly facilitates the computation of models for stars of low mass. At the center of the star

$$T_c = 1.246 \times 10^7 \mu \frac{M}{R} \Lambda(y) , \quad (14)$$

$$\rho_c = 8.446 \frac{M}{R^3} , \quad (15)$$

$$P_c = 8.680 \times 10^{15} \frac{M^2}{R^4} . \quad (16)$$

Therefore, degeneracy of the material does not affect the central density and the central pressure while it introduces a multiplicative factor in the expression for the central temperature T_c. In order to determine T_c, we must know $\Lambda(y)$ as a function of M, R, and μ. From equation 3 and the definition of N_e, we obtain

$$y = 6.185 \times 10^7 (1 + X) \frac{\rho}{T^{3/2}} . \quad (17)$$

Eliminating T and ρ between equations 17, 11, and 10, we obtain

$$\Lambda(y) = \frac{5.205 \times 10^{-2} (1 + X)^{2/3}}{y^{2/3} M^{1/3} \mu R} . \quad (18)$$

Thus knowing y, we can obtain $\Lambda(y)$ for a given M, R, and μ. The equations developed here have also been used by Limber (1958) in his study of the late type main sequence stars. However, we intend to apply this procedure not to the main sequence stars but the contracting stars of very low mass. We now apply these equations to compute the physical structure of the stars of very low mass.

COMPUTATION OF THE PHYSICAL STRUCTURE

We compute the physical structure for stars having the following two chemical compositions:

1. $X = 0.90$, $Y = 0.09$, $Z = 0.01$;

2. $X = 0.62$, $Y = 0.35$, $Z = 0.03$.

For each chemical composition and a given mass,

TABLE 1.—*The Relation Between y and D(y)*

y	D(y)	y	D(y)
0.000	1.0000	0.50	1.088
0.001	1.0002	0.60	1.105
0.005	1.0009	0.70	1.122
0.010	1.0017	0.80	1.140
0.015	1.0023	0.90	1.157
0.020	1.0035	1.00	1.174
0.025	1.0043	2.00	1.341
0.030	1.0053	3.00	1.504
0.035	1.0061	4.00	1.661
0.040	1.0070	5.00	1.814
0.045	1.0078	6.00	1.962
0.05	1.009	7.00	2.107
0.06	1.011	8.00	2.248
0.07	1.012	9.00	2.386
0.08	1.014	10.00	2.521
0.09	1.016	15.00	3.158
0.10	1.018	20.00	3.744
0.20	1.035	30.00	4.808
0.30	1.053	40.00	5.772
0.40	1.070		

we compute P_c, T_c, ρ_c, $\Lambda(y)$, and y at several values of the radius. The relation between y and $D(y)$ which is needed for these computations is given in table 1. This table has been prepared by making use of the tables of the Fermi-Dirac functions computed by McDougall and Stoner (1938). For the first composition, tables 2 to 7 give the physical quantities for M=0.09, 0.08, 0.07, 0.06, 0.05, 0.04, and at several radii for each mass. No models have been computed for those radii at which electron conduction becomes very important, and consequently the assumption of convective equilibrium does not hold. For each mass, the computation of models was stopped when the degeneracy parameter y reached a value close to 15. When y reaches this value, the material becomes appreciably degenerate and only then does the electron conduction become an efficient process for heat transport. However, to obtain a rough estimate of the central density and central temperature at a smaller radius, one more model was computed for each mass by

TABLE 2.—*Physical Properties of the Models*

$X=0.90$ $Y=0.09$ $Z=0.01$ Mass=0.09 $\mu=0.534$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
5.000	0.0173	1.0000	6.081E 03	1.198E 05	1.125E 11
4.000	0.0238	1.0000	1.188E 02	1.497E 05	2.746E 11
3.000	0.0375	1.0000	2.815E 02	1.996E 05	8.680E 11
2.000	0.0690	1.0000	9.502E 02	2.994E 05	4.394E 12
1.500	0.1060	1.0000	2.252E 01	3.993E 05	1.389E 13
1.000	0.1980	0.9823	7.601E 01	5.883E 05	7.031E 13
0.900	0.2335	0.9778	1.043E 00	6.506E 05	1.072E 14
0.800	0.2800	0.9745	1.485E 00	7.295E 05	1.717E 14
0.700	0.3450	0.9691	2.216E 00	8.291E 05	2.928E 14
0.600	0.4380	0.9645	3.519E 00	9.627E 05	5.425E 14
0.500	0.5840	0.9552	6.081E 00	1.144E 06	1.125E 15
0.400	0.8500	0.9298	1.188E 01	1.392E 06	2.746E 15
0.300	1.3750	0.8989	2.815E 01	1.794E 06	8.680E 15
0.200	3.0900	0.7972	9.502E 01	2.387E 06	4.394E 16
0.190	3.3900	0.7780	1.108E 02	2.452E 06	5.395E 16
0.180	3.8300	0.7574	1.303E 02	2.520E 06	6.698E 16
0.170	4.3900	0.7322	1.547E 02	2.579E 06	8.418E 16
0.160	5.0800	0.7059	1.856E 02	2.642E 06	1.073E 17
0.150	6.1000	0.6662	2.252E 02	2.660E 06	1.389E 17
0.140	7.3700	0.6295	2.770E 02	2.693E 06	1.830E 17
0.130	9.3500	0.5784	3.460E 02	2.665E 06	2.462E 17
0.120	12.5000	0.5163	4.399E 02	2.577E 06	3.391E 17
0.115	15.0000	0.4770	4.998E 02	2.484E 06	4.020E 17
0.100	32.2000	0.3288	7.601E 02	1.969E 06	7.031E 17

The limiting radius=0.0819

TABLE 3.—*Physical Properties of the Models*

$X=0.90$ $Y=0.09$ $Z=0.01$ Mass $=0.08$ $\mu=0.534$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
5.000	0.0184	1.0000	5.405E 03	1.065E 05	8.888E 10
4.000	0.0258	1.0000	1.056E 02	1.331E 05	2.170E 11
3.000	0.0396	1.0000	2.503E 02	1.774E 05	6.858E 11
2.000	0.0725	1.0000	8.446E 02	2.662E 05	3.472E 12
1.500	0.1130	1.0000	2.002E 01	3.549E 05	1.097E 13
1.000	0.2100	0.9822	6.757E 01	5.229E 05	5.555E 13
0.900	0.2480	0.9769	9.269E 01	5.778E 05	8.467E 13
0.800	0.2980	0.9724	1.320E 00	6.471E 05	1.356E 14
0.700	0.3670	0.9672	1.970E 00	7.355E 05	2.314E 14
0.600	0.4670	0.9611	3.128E 00	8.527E 05	4.296E 14
0.500	0.6200	0.9548	5.405E 00	1.017E 06	8.888E 14
0.400	0.9100	0.9240	1.056E 01	1.230E 06	2.170E 15
0.300	1.4700	0.8948	2.503E 01	1.588E 06	6.858E 15
0.200	3.2800	0.7858	8.446E 01	2.092E 06	3.472E 16
0.190	3.7100	0.7622	9.851E 01	2.136E 06	4.263E 16
0.180	4.2000	0.7407	1.159E 02	2.191E 06	5.292E 16
0.170	4.8100	0.7165	1.375E 02	2.244E 06	6.651E 16
0.160	5.6200	0.6862	1.650E 02	2.283E 06	8.477E 16
0.150	6.7500	0.6478	2.002E 02	2.299E 06	1.097E 17
0.140	8.3000	0.6048	2.462E 02	2.300E 06	1.446E 17
0.130	10.7500	0.5481	3.075E 02	2.244E 06	1.945E 17
0.120	14.8000	0.4798	3.910E 02	2.128E 06	2.679E 17
0.100	41.0000	0.2917	6.757E 02	1.553E 06	5.555E 17

The limiting radius $=0.0852$

making use of the condition of convective equilibrium. When y has a value of 30 or 40, electron conduction may destroy convective equilibrium in the central regions; still the convective model should give numerical results which have the correct order of magnitude.

For a given mass and chemical composition, there exists a limiting value of the radius below which there exists no model, for the material has become completely degenerate. At this stage, the star begins to approach closely a cooling curve in the H–R diagram. This limiting radius is obtained by using the asymptotic relation for $D(y)$:

$$D(y) = 0.4836\ y^{2/3} \quad (y \to \infty) . \qquad (19)$$

From equations 7 and 19, we have

$$\Lambda(y) = \frac{N_n + N_e}{N_n + 0.4836\ y^{2/3} N_e} \quad (y \to \infty) . \qquad (20)$$

It is now a straightforward procedure to obtain the limiting radius from equations 18 and 20. The limiting radii obtained in this way agree with those computed from the mass-radius relation for completely degenerate configurations. In figures 1 to 6, ρ_c and T_c have been plotted as a function of radius. (The figures follow the text.) As the radius of the star is decreased, the central density keeps on increasing while the temperature T_c shows an interesting variation. At first it increases and then reaches a maximum value. As the radius is further decreased, the temperature begins to decrease. This behavior of the temperature is due to the effects of degeneracy. We can visualize it physically in this way: When the radius changes from a large value to smaller ones, the star of a given mass can be pictured as a contracting star and only a part of the energy released by gravitational contraction is absorbed by the stellar material. However, when the material has become partially degenerate, the energy absorbed does not manifest itself as thermal energy. When partially degenerate gas is compressed, energy is needed to

TABLE 4.—*Physical Properties of the Models*

$X=0.90$ $Y=0.09$ $Z=0.01$ Mass $=0.07$ $\mu=0.534$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
4.000	0.0275	1.0000	9.238E 03	1.164E 05	1.661E 11
3.000	0.0423	1.0000	2.190E 02	1.553E 05	5.251E 11
2.000	0.0780	1.0000	7.390E 02	2.329E 05	2.658E 12
1.500	0.1210	0.9890	1.752E 01	3.071E 05	8.401E 12
1.000	0.2250	0.9811	5.912E 01	4.570E 05	4.253E 13
0.900	0.2660	0.9749	8.110E 01	5.046E 05	6.483E 13
0.800	0.3190	0.9715	1.155E 00	5.656E 05	1.038E 14
0.700	0.3940	0.9646	1.724E 00	6.419E 05	1.771E 14
0.600	0.5010	0.9588	2.737E 00	7.443E 05	3.282E 14
0.500	0.6780	0.9403	4.730E 00	8.760E 05	6.805E 14
0.400	0.9800	0.9195	9.238E 00	1.071E 06	1.661E 15
0.300	1.6000	0.8845	2.190E 01	1.373E 06	5.251E 15
0.200	3.5800	0.7752	7.390E 01	1.805E 06	2.658E 16
0.190	4.1000	0.7455	8.620E 01	1.828E 06	3.264E 16
0.180	4.6700	0.7215	1.014E 02	1.867E 06	4.052E 16
0.170	5.4000	0.6936	1.203E 02	1.900E 06	5.092E 16
0.160	6.3900	0.6585	1.443E 02	1.917E 06	6.490E 16
0.150	7.7000	0.6203	1.752E 02	1.926E 06	8.401E 16
0.140	9.6500	0.5718	2.155E 02	1.902E 06	1.107E 17
0.130	12.7500	0.5114	2.691E 02	1.832E 06	1.489E 17
0.125	15.0000	0.4759	3.027E 02	1.773E 06	1.742E 17
0.115	23.4000	0.3768	3.887E 02	1.526E 06	2.432E 17

The limiting radius $=0.0891$

bring the degenerate electrons closer and therefore the central temperature remains constant for a while and later it begins to decrease as a result of further contraction.

For the central temperature, the solid curve represents those models in which electron conduction is supposed to be negligible. We have assumed that as long as y remains equal to or less than 15 the stars under study can be represented accurately by completely convective models. The dotted part of the R–T_c curve represents models for which y is greater than 15 and therefore this section of the curve is not too accurate. The solid vertical line in each graph gives the limiting radius for a given mass. Table 8 gives the values of radius R, T_c, and ρ_c corresponding to the stage of maximum central temperature. The values of ρ_c and R for the stage of complete degeneracy are given in table 9.

In order to show the difference between the convective models composed of partially degenerate matter and of perfect gas we have computed the variation of temperature inside a star of mass 0.07 and radius 0.5. In table 10 are given the temperature distributions for the two cases. It also gives the density distribution and pressure distribution which are the same for both cases. In figure 7, these results are shown graphically where χ, the distance from the center of the star, is defined by $\chi=0.2737\xi$.

In tables 11 to 16, physically quantities are tabulated for the second chemical composition and figures 1 to 6 show the variation of T_c and ρ_c as a function of radius. Since the central density is independent of the chemical composition, the R–ρ_c curve for a given mass is the same for the two compositions considered here. The broken curve in each of these graphs shows the R–T_c relation for the second composition. The dotted part of this curve has the same significance as the dotted part of the curve for the first composition. The broken vertical line shows the limiting radius

TABLE 5.—*Physical Properties of the Models*

$X=0.90$ $Y=0.09$ $Z=0.01$ Mass $=0.06$ $\mu=0.534$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
3.000	0.0460	1.0000	1.877E 02	1.331E 05	3.858E 11
2.000	0.0844	1.0000	6.334E 02	1.996E 05	1.953E 12
1.500	0.1310	0.9876	1.502E 01	2.629E 05	6.172E 12
1.000	0.2430	0.9810	5.068E 01	3.917E 05	3.125E 13
0.900	0.2875	0.9745	6.951E 01	4.323E 05	4.763E 13
0.800	0.3470	0.9670	9.898E 01	4.826E 05	7.629E 13
0.700	0.4270	0.9624	1.477E 00	5.498E 05	1.301E 14
0.600	0.5470	0.9520	2.346E 00	6.335E 05	2.411E 14
0.500	0.7300	0.9424	4.054E 00	7.525E 05	5.000E 14
0.400	1.0600	0.9183	7.918E 00	9.166E 05	1.221E 15
0.300	1.7500	0.8767	1.877E 01	1.167E 06	3.858E 15
0.210	3.6700	0.7646	5.472E 01	1.454E 06	1.607E 16
0.200	4.0800	0.7481	6.334E 01	1.493E 06	1.953E 16
0.190	4.6400	0.7299	7.388E 01	1.519E 06	2.398E 16
0.180	5.3100	0.6971	8.689E 01	1.546E 06	2.977E 16
0.170	6.2100	0.6650	1.031E 02	1.562E 06	3.741E 16
0.160	7.4000	0.6289	1.237E 02	1.569E 06	4.768E 16
0.150	9.1000	0.5843	1.502E 02	1.555E 06	6.172E 16
0.140	11.7000	0.5295	1.847E 02	1.510E 06	8.134E 16
0.135	13.7000	0.4956	2.060E 02	1.466E 06	9.408E 16
0.120	24.7000	0.3754	2.933E 02	1.249E 06	1.507E 17

The limiting radius $=0.0938$

TABTE 6.—*Physical Properties of the Models*

$X=0.90$ $Y=0.09$ $Z=0.01$ Mass $=0.05$ $\mu=0.534$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
3.000	0.0500	1.0000	1.564E 02	1.109E 05	2.479E 11
2.000	0.0925	1.0000	5.279E 02	1.664E 05	1.356E 12
1.500	0.1440	0.9851	1.251E 01	2.185E 05	4.286E 12
1.000	0.2670	0.9790	4.223E 01	3.257E 05	2.170E 13
0.900	0.3160	0.9722	5.793E 01	3.594E 05	3.307E 13
0.800	0.3810	0.9653	8.248E 01	4.015E 05	5.298E 13
0.700	0.4690	0.9607	1.231E 00	4.566E 05	9.038E 13
0.600	0.6100	0.9406	1.955E 00	5.216E 05	1.674E 14
0.500	0.8100	0.9344	3.378E 00	6.218E 05	3.472E 14
0.400	1.1700	0.9144	6.598E 00	7.606E 05	8.477E 14
0.300	2.0000	0.8526	1.564E 01	9.456E 05	2.679E 15
0.210	4.2100	0.7145	4.560E 01	1.175E 06	1.116E 16
0.200	4.7800	0.7148	5.279E 01	1.189E 06	1.356E 16
0.190	5.5000	0.6871	6.157E 01	1.203E 06	1.665E 16
0.180	6.2000	0.6681	7.241E 01	1.235E 06	2.067E 16
0.170	7.4000	0.6289	8.596E 01	1.231E 06	2.598E 16
0.160	8.9000	0.5900	1.031E 02	1.227E 06	3.311E 16
0.150	11.3000	0.5373	1.251E 02	1.192E 06	4.286E 16
0.140	15.0000	0.4754	1.539E 02	1.130E 06	5.649E 16
0.120	35.0000	0.3162	2.444E 02	8.767E 05	1.046E 17

The limiting radius $=0.0997$

TABLE 7.—*Physical Properties of the Models*

$X=0.90$ $Y=0.09$ $Z=0.01$ Mass = 0.04 $\mu=0.534$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
2.000	0.1040	0.9882	4.223E 02	1.315E 05	8.680E 11
1.500	0.1610	0.9848	1.001E 01	1.747E 05	2.743E 12
1.000	0.3000	0.9759	3.378E 01	2.598E 05	1.389E 13
0.900	0.3550	0.9691	4.634E 01	2.866E 05	2.117E 13
0.800	0.4270	0.9640	6.598E 01	3.207E 05	3.391E 13
0.700	0.5290	0.9551	9.850E 01	3.632E 05	5.784E 13
0.600	0.6830	0.9397	1.564E 00	4.169E 05	1.072E 14
0.500	0.9100	0.9313	2.703E 00	4.958E 05	2.222E 14
0.400	1.3400	0.8996	5.279E 00	5.986E 05	5.425E 14
0.300	2.3000	0.8370	1.251E 01	7.426E 05	1.715E 15
0.250	3.3700	0.7780	2.162E 01	8.283E 05	3.555E 15
0.240	3.6900	0.7630	2.444E 01	8.462E 05	4.186E 15
0.230	4.0600	0.7470	2.777E 01	8.645E 05	4.963E 15
0.220	4.5000	0.7293	3.173E 01	8.823E 05	5.929E 15
0.210	5.0500	0.7072	3.648E 01	8.964E 05	7.141E 15
0.200	5.8000	0.6747	4.223E 01	8.979E 05	8.680E 15
0.190	6.6500	0.6510	4.925E 01	9.120E 05	1.066E 16
0.180	7.8000	0.6176	5.793E 01	9.133E 05	1.323E 16
0.170	9.4000	0.5775	6.876E 01	9.042E 05	1.663E 16
0.160	11.7000	0.5303	8.248E 01	8.822E 05	2.119E 16
0.155	13.3000	0.5027	9.072E 01	8.632E 05	2.406E 16
0.140	22.2000	0.3965	1.231E 02	7.538E 05	3.615E 16

The limiting radius = 0.1074

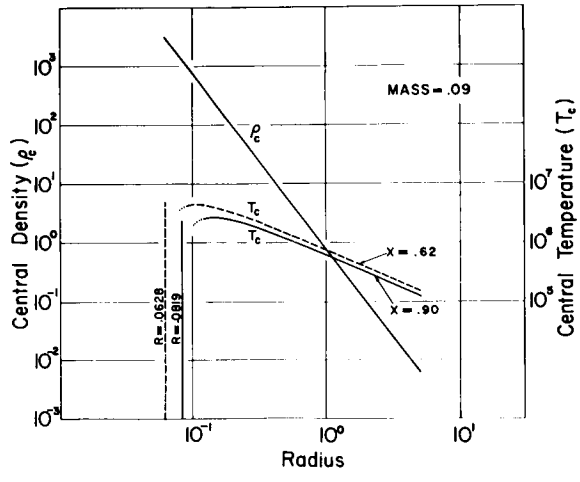


FIGURE 1.—The central temperature and the central density in completely convective models. Mass = 0.09.

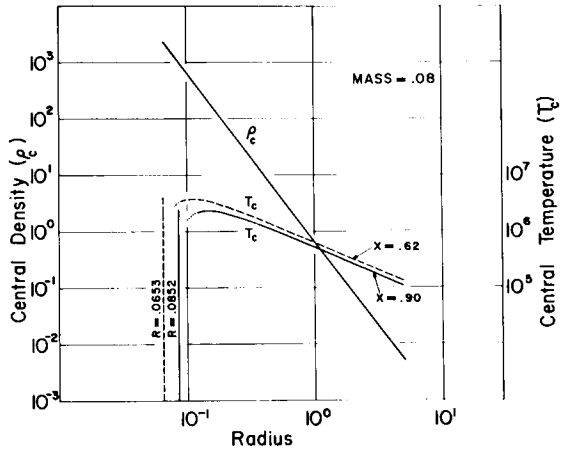


FIGURE 2.—The central temperature and the central density in completely convective models. Mass = 0.08.

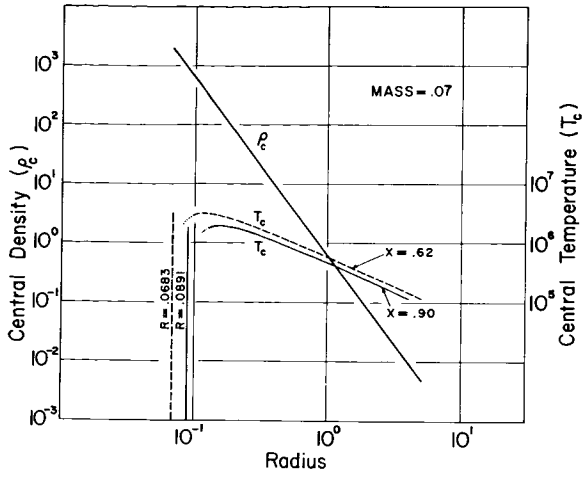


FIGURE 3.—The central temperature and the central density in completely convective models. Mass = 0.07.

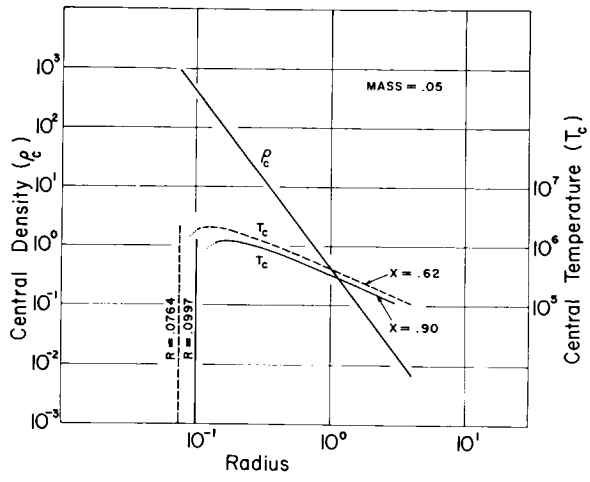


FIGURE 5.—The central temperature and the central denfity in completely convective models. Mass = 0.05.

for a model of a given mass having the second composition. For this composition the properties of the stars at the stage of maximum central temperature are given in table 8, whereas those corresponding to the stage of complete degeneracy are given in table 9. The temperature and density distribution for a star of mass 0.07 and radius 0.5 for this chemical composition are given in table 17 and they are shown graphically in figure 8. It is interesting to plot the central density—central temperature relation for a given mass in the temperature-density diagram in which the degenerate region can be separated from the nondegenerate one. For the first composition figure 9 shows the positions of various models in such a diagram. For a given mass, the $\log\rho_c$–$\log T_c$ gives an evolutionary path for a contracting star. Figure 10 shows the same graph for the second chemical composition. The boundary between the nondegenerate and degenerate regions is obtained by equating the electron pressure from the perfect gas law to that obtained from the equation of state for a completely degenerate gas.

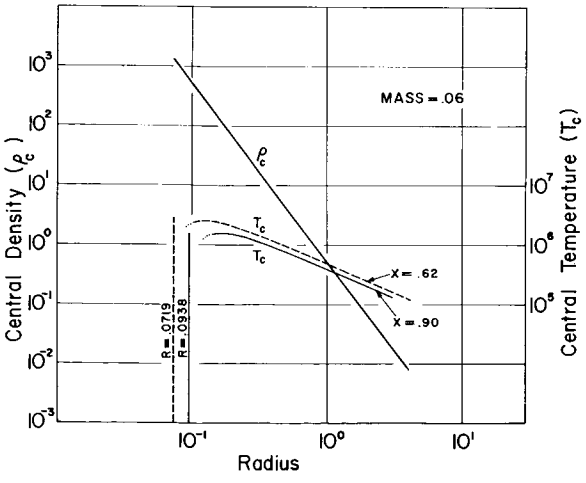


FIGURE 4.—The central temperature and the central density in completely convective models. Mass = 0.06.

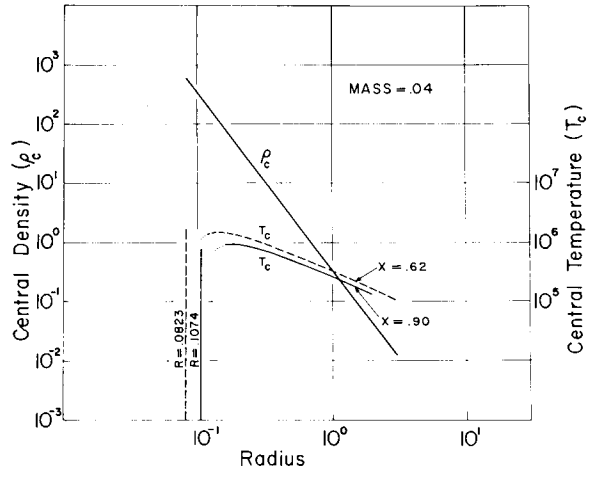


FIGURE 6.—The central temperature and the central density in completely convective models. Mass = 0.04.

TABLE 8.—*Properties at Maximum Central Temperature*

Mass	X = 0.90 Y = 0.09 Z = 0.01			X = 0.62 Y = 0.35 Z = 0.03		
	Radius	Temperature	Density	Radius	Temperature	Density
0.09	0.14	2.693×10^6	2.770×10^2	0.11	4.420×10^6	5.711×10^2
0.08	0.14	2.300×10^6	2.462×10^2	0.11	3.761×10^6	5.076×10^2
0.07	0.15	1.926×10^6	1.752×10^2	0.12	3.151×10^6	3.421×10^2
0.06	0.16	1.569×10^6	1.237×10^2	0.12	2.574×10^6	2.933×10^2
0.05	0.18	1.235×10^6	7.241×10^1	0.13	2.018×10^6	1.922×10^2
0.04	0.18	9.133×10^5	5.793×10^1	0.14	1.500×10^6	1.231×10^2

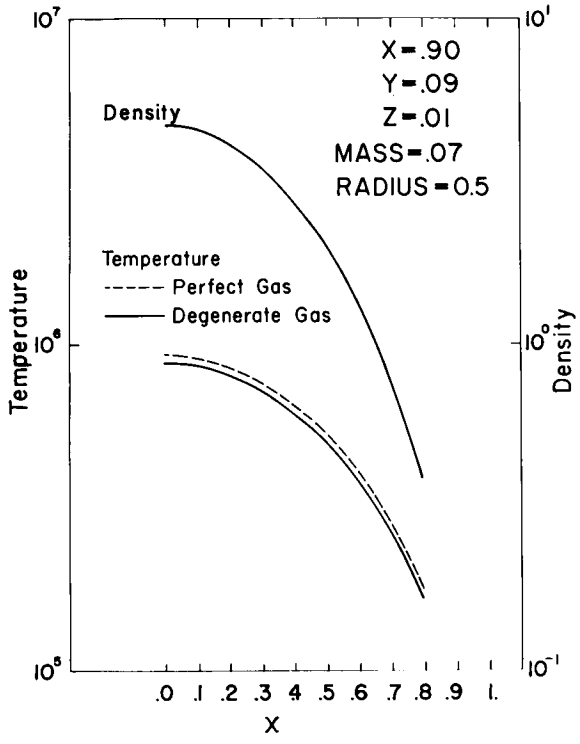


FIGURE 7.—The temperature and density distributions in a completely convective model. Population II.

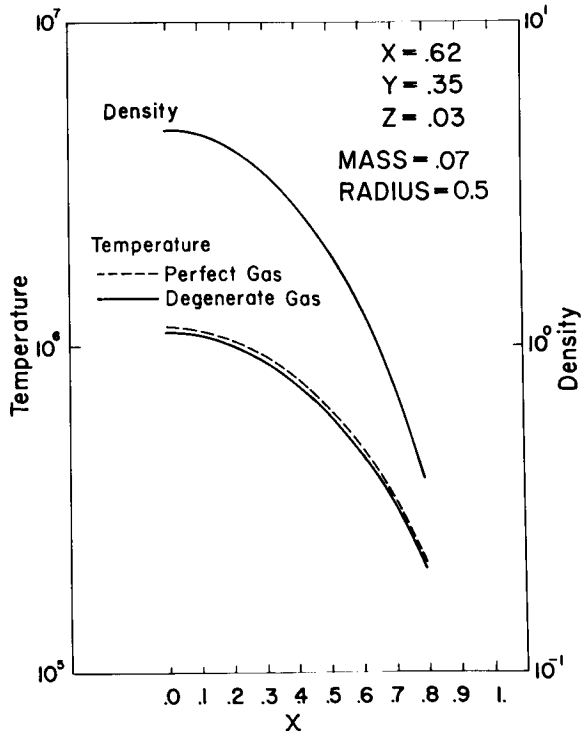


FIGURE 8.—The temperature and density distributions in a completely convective model. Population I.

TABLE 9.—*Properties of the Completely Degenerate Models*

Mass	X = 0.90 Y = 0.09 Z = 0.01		X = 0.62 Y = 0.35 Z = 0.03	
	Limiting Radius	Limiting Density	Limiting Radius	Limiting Density
0.09	0.0819	1.384×10^3	0.0628	3.069×10^3
0.08	0.0852	1.092×10^3	0.0653	2.427×10^3
0.07	0.0891	8.357×10^2	0.0683	1.856×10^3
0.06	0.0938	6.141×10^2	0.0719	1.363×10^3
0.05	0.0997	4.261×10^2	0.0764	9.471×10^2
0.04	0.1074	2.726×10^2	0.0823	6.059×10^2

TABLE 10.—*Temperature, Density, and Pressure Distributions*

X = 0.90 Y = 0.09 Z = 0.01 μ = 0.534 Mass = 0.07 Radius = 0.5

$\frac{r}{R}$	Temperature		Density	Pressure
	Perfect Gas	Degenerate Gas		
0.0000	9.316×10^5	8.760×10^5	4.730	6.805×10^{14}
0.0547	9.255×10^5	8.702×10^5	4.683	6.693×10^{14}
0.1095	9.071×10^5	8.530×10^5	4.545	6.366×10^{14}
0.1642	8.772×10^5	8.248×10^5	4.322	5.430×10^{14}
0.2190	8.369×10^5	7.869×10^5	4.027	5.204×10^{14}
0.2737	7.874×10^5	7.404×10^5	3.675	4.469×10^{14}
0.3284	7.304×10^5	6.868×10^5	3.284	3.703×10^{14}
0.3832	6.676×10^5	6.277×10^5	2.869	2.958×10^{14}
0.4379	6.007×10^5	5.648×10^5	2.449	2.272×10^{14}
0.4926	5.317×10^5	4.999×10^5	2.039	1.674×10^{14}
0.5474	4.620×10^5	4.344×10^5	1.675	1.179×10^{14}
0.6295	3.595×10^5	3.380×10^5	1.134	6.295×10^{13}
0.6842	2.943×10^5	2.767×10^5	8.400×10^{-1}	3.817×10^{13}
0.7390	2.326×10^5	2.187×10^5	5.903×10^{-1}	2.120×10^{13}
0.7937	1.750×10^5	1.646×10^5	3.853×10^{-1}	1.103×10^{13}
1.0000	0.000	0.000	0.000	0.000

TABLE 11.—*Physical Properties of the Models*

$X=0.62$ $Y=0.35$ $Z=0.03$ Mass = 0.09 $\mu=0.659$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
5.000	0.0105	1.0000	6.081E 03	1.478E 05	1.125E 11
4.000	0.0150	1.0000	1.188E 02	1.847E 05	2.746E 11
3.000	0.0231	1.0000	2.815E 02	2.463E 05	8.680E 11
2.000	0.0427	1.0000	9.502E 02	3.695E 05	4.394E 12
1.500	0.0660	1.0000	2.252E 01	4.927E 05	1.389E 13
1.000	0.1218	1.0000	7.601E 01	7.390E 05	7.031E 13
0.900	0.1430	0.9880	1.043E 00	8.112E 05	1.072E 14
0.800	0.1715	0.9851	1.485E 00	9.100E 05	1.717E 14
0.700	0.2110	0.9802	2.216E 00	1.035E 06	2.928E 14
0.600	0.2665	0.9787	3.519E 00	1.205E 06	5.425E 14
0.500	0.3550	0.9701	6.081E 00	1.434E 06	1.125E 15
0.400	0.5080	0.9548	1.188E 01	1.764E 06	2.746E 15
0.300	0.8180	0.9267	2.815E 01	2.283E 06	8.680E 15
0.200	1.6600	0.8673	9.502E 01	3.205E 06	4.394E 16
0.150	2.9000	0.7970	2.252E 02	3.926E 06	1.389E 17
0.140	3.3900	0.7696	2.770E 02	4.062E 06	1.830E 17
0.130	4.0200	0.7366	3.460E 02	4.187E 06	2.462E 17
0.120	4.8900	0.7035	4.399E 02	4.332E 06	3.391E 17
0.110	6.2000	0.6580	5.711E 02	4.420E 06	4.802E 17
0.100	8.3000	0.5932	7.601E 02	4.384E 06	7.031E 17
0.098	8.9300	0.5770	8.076E 02	4.352E 06	7.623E 17
0.096	9.6000	0.5616	8.592E 02	4.323E 06	8.278E 17
0.090	12.2500	0.5089	1.043E 03	4.179E 06	1.072E 18
0.080	22.5000	0.3810	1.485E 03	3.519E 06	1.717E 18

The limiting radius = 0.0628

TABLE 12.—*Physical Properties of the Models*

$X=0.62$ $Y=0.35$ $Z=0.03$ Mass $=0.08$ $\mu=0.659$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
5.000	0.0115	1.0000	5.405E 03	1.314E 05	8.888E 10
4.000	0.0160	1.0000	1.056E 02	1.642E 05	2.170E 11
3.000	0.0244	1.0000	2.503E 02	2.190E 05	6.858E 11
2.000	0.0453	1.0000	8.446E 02	3.284E 05	3.472E 12
1.500	0.0700	1.0000	2.002E 01	4.379E 05	1.097E 13
1.000	0.1295	1.0000	6.757E 01	6.569E 05	5.555E 13
0.900	0.1520	0.9867	9.269E 01	7.201E 05	8.467E 13
0.800	0.1820	0.9841	1.320E 00	8.080E 05	1.356E 14
0.700	0.2240	0.9797	1.970E 00	9.193E 05	2.314E 14
0.600	0.2840	0.9755	3.128E 00	1.068E 06	4.286E 14
0.500	0.3780	0.9675	5.405E 00	1.271E 06	8.888E 14
0.400	0.5400	0.9536	1.056E 01	1.566E 06	2.170E 15
0.300	0.8700	0.9251	2.503E 01	2.026E 06	6.858E 15
0.200	1.7700	0.8647	8.446E 01	2.840E 06	3.472E 16
0.150	3.1800	0.7806	2.002E 02	3.418E 06	1.097E 17
0.130	4.4000	0.7257	3.075E 02	3.667E 06	1.945E 17
0.120	5.4500	0.6800	3.910E 02	3.722E 06	2.679E 06
0.110	6.9800	0.6299	5.076E 02	3.761E 06	3.794E 17
0.100	9.5500	0.5620	6.757E 02	3.692E 06	5.555E 17
0.099	9.9000	0.5542	6.964E 02	3.677E 06	5.783E 17
0.098	10.2500	0.5491	7.179E 02	3.680E 06	6.023E 17
0.090	14.6000	0.4691	9.269E 02	3.424E 06	8.467E 17
0.080	28.9000	0.3363	1.320E 03	2.761E 06	1.356E 18

The limiting radius $=0.0653$

TABLE 13.—*Physical Properties of the Models*

$X=0.62$ $Y=0.35$ $Z=0.03$ Mass = 0.07 $\mu=0.659$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
5.000	0.0125	1.0000	4.730E 03	1.150E 05	6.805E 10
4.000	0.0170	1.0000	9.238E 03	1.437E 05	1.661E 11
3.000	0.0262	1.0000	2.190E 02	1.916E 05	5.251E 11
2.000	0.0483	1.0000	7.390E 02	2.874E 05	2.658E 12
1.500	0.0748	1.0000	1.752E 01	3.832E 05	8.401E 12
1.000	0.1382	0.9892	5.912E 01	5.686E 05	4.253E 13
0.900	0.1628	0.9852	8.110E 01	6.292E 05	6.483E 13
0.800	0.1950	0.9828	1.155E 00	7.061E 05	1.038E 14
0.700	0.2400	0.9780	1.724E 00	8.030E 05	1.771E 14
0.600	0.3050	0.9726	2.737E 00	9.317E 05	3.282E 14
0.500	0.4050	0.9660	4.730E 00	1.110E 06	6.805E 14
0.400	0.5810	0.9493	9.238E 00	1.364E 06	1.661E 15
0.300	0.9350	0.9218	2.190E 01	1.766E 06	5.251E 15
0.200	1.9400	0.8502	7.390E 01	2.443E 06	2.658E 16
0.150	3.5000	0.7649	1.752E 02	2.931E 06	8.401E 16
0.140	4.1000	0.7379	2.155E 02	3.029E 06	1.107E 17
0.130	4.9400	0.7014	2.691E 02	3.101E 06	1.489E 17
0.120	6.1500	0.6579	3.421E 02	3.151E 06	2.051E 17
0.110	8.0500	0.5980	4.442E 02	3.125E 06	2.905E 17
0.100	11.2000	0.5288	5.912E 02	3.039E 06	4.253E 17
0.098	12.2500	0.5091	6.282E 02	2.986E 06	4.611E 17
0.095	14.2000	0.4741	6.896E 02	2.868E 06	5.222E 17
0.080	40.0000	0.2825	1.155E 03	2.030E 06	1.038E 18

The limiting radius = 0.0683

TABLE 14.—*Physical Properties of the Models*

$X=0.62$ $Y=0.35$ $Z=0.03$ Mass $=0.06$ $\mu=0.659$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
4.000	0.0186	1.0000	7.918E 03	1.232E 05	1.221E 11
3.000	0.0284	1.0000	1.877E 02	1.642E 05	3.858E 11
2.000	0.0521	1.0000	6.334E 02	2.463E 05	1.953E 12
1.500	0.0810	1.0000	1.502E 01	3.284E 05	6.172E 12
1.000	0.1500	0.9862	5.068E 01	4.859E 05	3.125E 13
0.900	0.1760	0.9847	6.951E 01	5.390E 05	4.763E 13
0.800	0.2110	0.9819	9.898E 01	6.047E 05	7.629E 13
0.700	0.2600	0.9762	1.477E 00	6.870E 05	1.301E 14
0.600	0.3300	0.9715	2.346E 00	7.977E 05	2.411E 14
0.500	0.4400	0.9623	4.054E 00	9.482E 05	5.000E 14
0.400	0.6330	0.9440	7.918E 00	1.163E 06	1.221E 15
0.300	1.0250	0.9053	1.877E 01	1.487E 06	3.858E 15
0.200	2.1500	0.8355	6.334E 01	2.058E 06	1.953E 16
0.180	2.6400	0.8094	8.689E 01	2.215E 06	2.977E 16
0.150	3.9300	0.7424	1.502E 02	2.438E 06	6.172E 16
0.140	4.6800	0.7100	1.847E 02	2.498E 06	8.134E 16
0.130	5.7000	0.6712	2.307E 02	2.544E 06	1.094E 17
0.120	7.1800	0.6270	2.933E 02	2.574E 06	1.507E 17
0.110	9.5500	0.5624	3.807E 02	2.519E 06	2.134E 17
0.100	14.0000	0.4800	5.068E 02	2.365E 06	3.125E 17
0.090	24.8000	0.3639	6.951E 02	1.992E 06	4.763E 17

The limiting radius $=0.0719$

TABLE 15.—*Physical Properties of the Models*

$X=0.62$ $Y=0.35$ $Z=0.03$ Mass = 0.05 $\mu=0.659$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
4.000	0.0202	1.0000	6.598E 03	1.026E 05	8.477E 10
3.000	0.0310	1.0000	1.564E 02	1.368E 05	2.679E 11
2.000	0.0575	1.0000	5.279E 02	2.053E 05	1.356E 12
1.500	0.0886	1.0000	1.251E 01	2.737E 05	4.286E 12
1.000	0.1645	0.9854	4.223E 01	4.045E 05	2.170E 13
0.900	0.1930	0.9841	5.793E 01	4.489E 05	3.307E 13
0.800	0.2315	0.9807	8.248E 01	5.033E 05	5.298E 13
0.700	0.2850	0.9758	1.231E 00	5.723E 05	9.038E 13
0.600	0.3630	0.9688	1.955E 00	6.629E 05	1.674E 14
0.500	0.4850	0.9584	3.378E 00	7.869E 05	3.472E 14
0.400	0.7080	0.9309	6.598E 00	9.554E 05	8.477E 14
0.300	1.1400	0.9038	1.564E 01	1.237E 06	2.679E 15
0.200	2.4000	0.8249	5.279E 01	1.693E 06	1.356E 16
0.180	3.0000	0.7899	7.241E 01	1.802E 06	2.067E 16
0.160	3.9100	0.7426	1.031E 02	1.905E 06	3.311E 16
0.150	4.5500	0.7171	1.251E 02	1.963E 06	4.286E 16
0.140	5.5000	0.6794	1.539E 02	1.992E 06	5.649E 16
0.130	6.7500	0.6390	1.922E 02	2.018E 06	7.598E 16
0.120	8.7000	0.5814	2.444E 02	1.989E 06	1.046E 17
0.110	12.0000	0.5132	3.173E 02	1.915E 06	1.482E 07
0.090	38.0000	0.2909	5.793E 02	1.327E 06	3.307E 17

The limiting radius = 0.0764

TABLE 16.—*Physical Properties of the Models*

$X=0.62$ $Y=0.35$ $Z=0.03$ Mass $=0.04$ $\mu=0.659$

Radius	Y	Lambda (Y)	Density (C)	Temperature (C)	Pressure (C)
3.000	0.0350	1.0000	1.251E 02	1.095E 05	1.715E 11
2.000	0.0642	1.0000	4.223E 02	1.642E 05	8.680E 11
1.500	0.0995	0.9893	1.001E 01	2.166E 05	2.743E 12
1.000	0.1840	0.9852	3.378E 01	3.236E 05	1.389E 13
0.900	0.2160	0.9833	4.634E 01	3.588E 05	2.117E 13
0.800	0.2590	0.9803	6.598E 01	4.025E 05	3.391E 13
0.700	0.3200	0.9731	9.850E 01	4.566E 05	5.784E 13
0.600	0.4090	0.9639	1.564E 00	5.276E 05	1.072E 14
0.500	0.5460	0.9540	2.703E 00	6.267E 05	2.222E 14
0.400	0.7930	0.9298	5.279E 00	7.634E 05	5.425E 14
0.300	1.3000	0.8917	1.251E 01	9.762E 05	1.715E 15
0.250	1.8000	0.8615	2.162E 01	1.132E 06	3.555E 15
0.200	2.8000	0.8017	4.223E 01	1.317E 06	8.680E 15
0.190	3.1300	0.7836	4.925E 01	1.355E 06	1.066E 16
0.180	3.5400	0.7629	5.793E 01	1.392E 06	1.323E 16
0.170	4.0300	0.7378	6.876E 01	1.425E 06	1.663E 16
0.160	4.6900	0.7114	8.248E 01	1.460E 06	2.119E 16
0.150	5.5500	0.6764	1.001E 02	1.481E 06	2.734E 16
0.140	6.7500	0.6393	1.231E 02	1.500E 06	3.615E 16
0.130	8.5900	0.5836	1.538E 02	1.474E 06	4.863E 16
0.120	11.4500	0.5216	1.955E 02	1.428E 06	6.698E 16
0.100	30.5000	0.3187	3.378E 02	1.047E 06	1.389E 17

The limiting radius $=0.0823$

TABLE 17.—*Temperature, Density and Pressure Distributions*

X = 0.62 Y = 0.35 Z = 0.03 μ = 0.659 Mass = 0.07 Radius = 0.5

$\frac{r}{R}$	Temperature		Density	Pressure
	Perfect Gas	Degenerate Gas		
0.0000	1.149×10^{6}	1.110×10^{6}	4.730	6.805×10^{14}
0.0547	1.141×10^{6}	1.103×10^{6}	4.683	6.693×10^{14}
0.1095	1.119×10^{6}	1.081×10^{6}	4.545	6.366×10^{14}
0.1642	1.082×10^{6}	1.045×10^{6}	4.322	5.430×10^{14}
0.2190	1.032×10^{5}	9.971×10^{5}	4.027	5.204×10^{14}
0.2737	9.711×10^{5}	9.382×10^{5}	3.675	4.469×10^{14}
0.3284	9.008×10^{5}	8.702×10^{5}	3.284	3.703×10^{14}
0.3832	8.234×10^{5}	7.954×10^{5}	2.869	2.958×10^{14}
0.4379	7.409×10^{5}	7.157×10^{5}	2.449	2.272×10^{14}
0.4926	6.557×10^{5}	6.335×10^{5}	2.039	1.674×10^{14}
0.5474	5.698×10^{5}	5.504×10^{5}	1.675	1.179×10^{14}
0.6295	4.434×10^{5}	4.283×10^{5}	1.134	6.295×10^{13}
0.6842	3.630×10^{5}	3.506×10^{5}	8.400×10^{-1}	3.817×10^{13}
0.7390	2.869×10^{5}	2.772×10^{5}	5.903×10^{-1}	2.120×10^{13}
0.7937	2.159×10^{5}	2.086×10^{5}	3.853×10^{-1}	1.103×10^{13}
1.0000	0.000	0.000	0.000	0.000

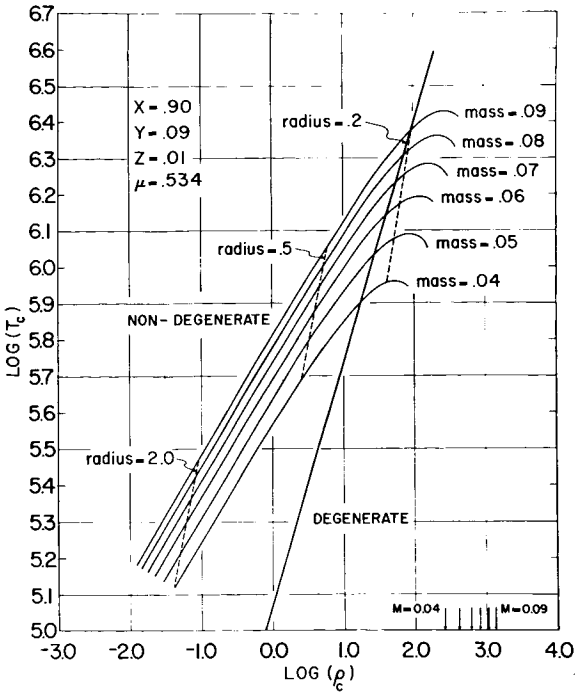


FIGURE 9.—The temperature-density diagram for completely convective models. Population II.

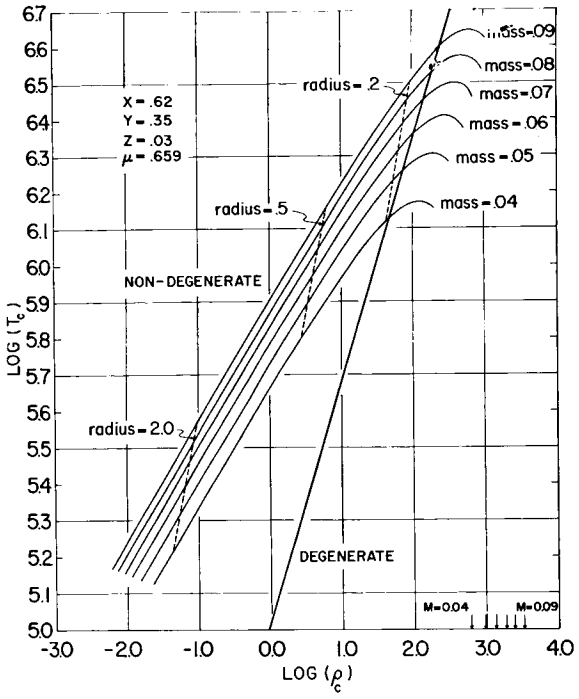


FIGURE 10.—The temperature-density diagram for completely convective models. Population I.

DISCUSSION

The numerical results presented here show clearly that for a given composition, there exists a limiting mass below which a contracting star cannot reach the main sequence stage, because the temperature and density at the center are too low for hydrogen burning to start. Instead the star becomes a degenerate star as a consequence of the contraction. After the star has evolved beyond the stage of maximum central temperature, further contraction will take it towards the stage of complete degeneracy. Thus, all stars having a mass less than a certain limiting mass ultimately become completely degenerate objects without ever going through the normal stellar evolution. The exact determination of this limiting mass for a given composition requires a knowledge of the luminosity of the contracting stars which can be obtained if we know the atmospheric structure in addition to the interior models computed here. Suitable model atmospheres for contracting stars of low mass are being computed which, together with the interior models presented here, will give us not only the evolutionary tracks in the H–R diagram for these stars, but also the limiting mass which gives a lower limit to the mass of a main sequence star and the time scale for the Helmholtz–Kelvin contraction. If we assume reasonable luminosities for these stars, we find that, for stars with population I composition, the limiting mass is approximately 0.07. Similarly for the population II stars the limiting mass is approximately 0.09.

ACKNOWLEDGMENTS

It is a pleasure to express my gratitude to Dr. A. G. W. Cameron and Dr. Edward Upton for their helpful discussions. Mr. Alan Liebman is thanked for his assistance in numerical computations.

REFERENCES

1. HAYASHI, C., *P.A.S.J.*, **13**, 450, 1962.
2. LIMBER, D. N., *Ap. J.*, **127**, 387, 1959.
3. MCDOUGALL, J., and STONER, E. C., *Phil. Trans. R. Soc. London*, **237**, 67, 1938.
4. TOLMAN, R. C., *The Principles of Statistical Mechanics* (London: Oxford University Press), pp. 388–390, 1938.

COSMIC-RAY-INDUCED STABLE AND RADIOACTIVE NUCLIDES IN METEORITES

MICHAEL E. LIPSCHUTZ*

Goddard Space Flight Center

Meteorites, the only extra-terrestrial solid objects currently available on the earth, are being extensively studied by many investigators. Data now available enable us to reconstruct some features of that period when meteorites were exposed to cosmic rays.

A brief discussion of meteoritic types is followed by presentation of the results of radiochemical and mass-spectrometric investigations on meteorites. From the amounts of cosmogenic stable and radioactive species present, conclusions can be drawn concerning the duration of cosmic ray exposure, the spectrum and intensity of the incident cosmic ray flux, the pre-atmospheric size and shape of the meteorites, and the terrestrial history of these objects. To close, a brief discussion of some, as yet unanswered, questions concerning meteorites is presented.

INTRODUCTION

In the few years since the first artifical earth satellite was launched, direct measurements of the interplanetary radiation environment have become relatively common. Such measurements can indicate only what is occurring here and now. They cannot, for example, tell us whether the cosmic ray intensity or energy spectra have varied in the last few million years or whether the intensity is the same near the sun as near Jupiter. Related questions, such as the erosion rate of solid objects by high velocity dust particles or the probability of an object's collision with large pieces of cosmic "debris," are also difficult to answer by use of artificial probes. Effects due to these processes are detectable only in objects that have been in space for very long periods.

Meteorites are both the oldest material and the only extra-terrestrial samples of which we know. They are therefore unique objects by which the early nucleosynthetic and space environmental conditions can be examined. It is the purpose of this paper to review some of the significant observations regarding interplanetary environments as deduced from the study of meteorites.

The principal classification of meteorites is on the basis of physical appearance. Those that are almost pure iron-niclek alloys are called *siderites*. Stony-iron meteorites (*siderolites*) are further subdivided into *pallasites* in which the stony constituents are embedded in a matrix of metal, and *mesosiderites*, in which bits of iron-nickel are in a stone matrix. Stony meteorites (*aerolites*) are divided into *chondrites*, containing spherical inclusions called chondrules, and *achondrites*. The number and proportion of each class[1] are listed in table 1. "Falls" refer to meteorites that are seen to fall and are recovered, while "finds" are meteorites whose falls have not been witnessed. It seems probable that the falls more nearly represent the preterrestrial compositional distribution of meteorites than the finds, which would be biased in the direction of the more easily recognized irons. The average chemical composition[2] of both iron and stone meteorites is listed in table 2. In general, only a few elements are abundant enough to concern the scientest studying cosmic radiation effects; these elements include Fe, Ni, O, Si, and Mg.

*Dr. Lipschutz, on active duty with the U.S. Army, is stationed at Goddard Space Flight Center.

TABLE 1.—*Classification of Various Types of Meteorites* [1]

Types of meteorites	Falls	Finds	Percent falls	Percent finds
Irons (Siderites)	42	503	6.6	61.1
Stony-Irons (Siderolites):				
Pallasites	2	37	0.3	4.5
Mesosiderites	10	18	1.6	2.2
Stones (Aerolites):				
Chondrites	523	254	82.7	30.9
Achondrites	56	11	8.8	1.3
Total	633	823	100	100

Recovered meteorites range in mass from a few hundredths of a gram to 60 metric tons.[3] Larger meteorites have at times hit the earth: The object that created the "Great Arizona Meteorite Crater" weighed between 10^4 and 10^5 metric tons,[4] and the shock pressure generated by its impact was more than sufficient to convert graphite to diamond.[5] Brown[6,7] has reported on the in-fall rate of meteorites of various masses.

It is generally accepted from theoretical studies[8] that most, if not all, meteorites come from the asteroidal belt—a group of small planetsimals rotating about the sun in orbits between those of Mars and Jupiter. On the basis of chemical and mineralogical studies Urey[9] suggested that some meteorites come from the moon's surface. It is not certain that this is the case; the only recovered meteorite whose orbit was calculated[10] came from the asteroidal belt.

TABLE 2.—*Average Chemical Composition of Stony and Iron Meteorites*[2]

Element	Weight percent	
	Stone meteorites	Iron meteorites
O	35.71	----------
Fe	23.31	89.70
Si	18.07	----------
Mg	13.87	----------
S	1.80	0.08
Ca	1.73	----------
Ni	1.53	9.10
Al	1.52	----------
Na	0.65	----------
Cr	0.32	----------
K	0.17	----------
C	0.15	0.12
Co	0.12	0.62
P	0.11	0.18
Ti	0.11	----------

COSMIC-RAY-INDUCED NUCLEAR REACTIONS

While objects are in space, they are being continuously bombarded by high energy cosmic rays. Even the earth is subjected to this bombardment, but its thick atmosphere (1010 gm/cm^2) absorbs most of the particles and allows but a few to strike the surface.

Most studies of cosmic ray interaction with meteoritic material have employed mass-spectrometric and radiochemical techniques, and accordingly this same approach will be emphasized here. However, other means can be used in studying interplanetary environments. Houtermans and coworkers[11] have studied the thermoluminescent glow curves of some stony meteorites to measure their accumulated radiation dosage. Heymann and Fluit,[12] using 20-Kev A^+ atoms, have estimated the relative space erosion rates of stone and iron meteorites by studying their surface "sputtering." The application of other physical techniques to investigations of this type would indeed be of value.

Cosmic ray primaries are about 87% protons

and 13% alpha particles. Secondary particles that consist mostly of high and low energy protons, neutrons, and mesons are produced by the interaction of primaries and matter. Nuclear reactions that can be induced by such particles may be grouped into three categories: *spallation*, *neutron capture*, and *stripping*. In general, it may be said that any energetically possible nuclear reaction that *can* occur, will.

The first, and most important, of these reactions, *spallation*, occurs when a high energy particle strikes a nucleus and deposits enough energy to "evaporate" some nucleons. The resulting atom may be either radioactive or stable, and can differ in mass and atomic numbers by as few as 1 or 2 or as many as 40 to 50. Such a reaction may be represented by

$$ {}_{Z}M^{A}+p\rightarrow X(n+p)+{}_{a}N^{b}, \qquad (1) $$

where Z and A are, respectively, the atomic and mass numbers of the target nucleus M; n represents a neutron; p, a proton; and N, the product nucleus. X, a, and b are integers that, to a first approximation, are dependent on the energy of the impacting particle. Some of the n and p that are produced may be associated as nuclear fragments, such as heavy hydrogen or helium nuclei. Examples of Reaction 1 are:

$$ {}_{26}\mathrm{Fe}^{56}+{}_{1}p^{1}\rightarrow {}_{26}\mathrm{Fe}^{55}(t_{1/2}=2.6 \text{ years})+{}_{0}n^{1}+{}_{1}p^{1}, \qquad (1a) $$

$$ {}_{26}\mathrm{Fe}^{56}+{}_{1}p^{1}\rightarrow {}_{21}\mathrm{Sc}^{45} \text{ (stable)}+6\ {}_{1}p^{1}+6\ {}_{0}n^{1}, \qquad (1b) $$

$$ {}_{26}\mathrm{Fe}^{56}+{}_{1}p^{1}\rightarrow {}_{18}\mathrm{Al}^{26}(t_{1/2}=7.4\times 10^{5} \text{ years})+14\ {}_{1}p^{1}+17\ {}_{0}n^{1}. \qquad (1c) $$

Neutron capture occurs when a low energy neutron produced by a reaction of type 1 strikes a nucleus. It may be represented by:

$$ {}_{Z}M^{A}+{}_{0}n^{1}\rightarrow {}_{Z}M^{A+1}+\gamma. \qquad (2) $$

The product is one atomic number different from the target nucleus. Examples are:

$$ {}_{28}\mathrm{Ni}^{58}+{}_{0}n^{1}\rightarrow {}_{28}\mathrm{Ni}^{59}(t_{1/2}=8.0\times 10^{4} \text{ years})+\gamma, \qquad (2a) $$

$$ {}_{17}\mathrm{Cl}^{35}+{}_{0}n^{1}\rightarrow {}_{17}\mathrm{Cl}^{36}(t_{1/2}=3.08\times 10^{5} \text{ years})+\gamma. \qquad (2b) $$

The third type of interaction, the *stripping* reaction, occurs when an alpha particle more or less elastically strikes a nucleus. In this case a nucleon is stripped from the alpha particle, and the fragment continues on. These reactions may be represented by:

$$ {}_{a}M^{b}+{}_{2}\mathrm{He}^{4}\rightarrow {}_{1}H^{3}(t_{1/2}=12.4 \text{ years})+{}_{y}N^{z}+X(n+p), \qquad (3a) $$

$$ {}_{a}M^{b}+{}_{2}\mathrm{He}^{4}\rightarrow {}_{2}\mathrm{He}^{3} \text{ (stable)}+{}_{y}N^{z}+X(n+p). \qquad (3b) $$

Nuclei heavier than about mass 64 (nickel) are quite rare in meteorites, so that the amount of material produced by fission is negligible.

DATA FROM METEORITES

There recently have been several studies aimed at determining the primary and secondary cosmic ray spectra at various depths in iron meteorites. The calculation of the spectra is not at all "straightforward" because both primary and secondary particles can cause nuclear reactions. Goel,[13] on the basis of nuclear-star production in iron, has calculated the relative production rates of several specific nuclides in iron; the agreement between his calculations and the observed ratios is fair. Perhaps the most successful calculation to date has been that by Arnold et al.,[14] whose results agree well with the observed concentrations of

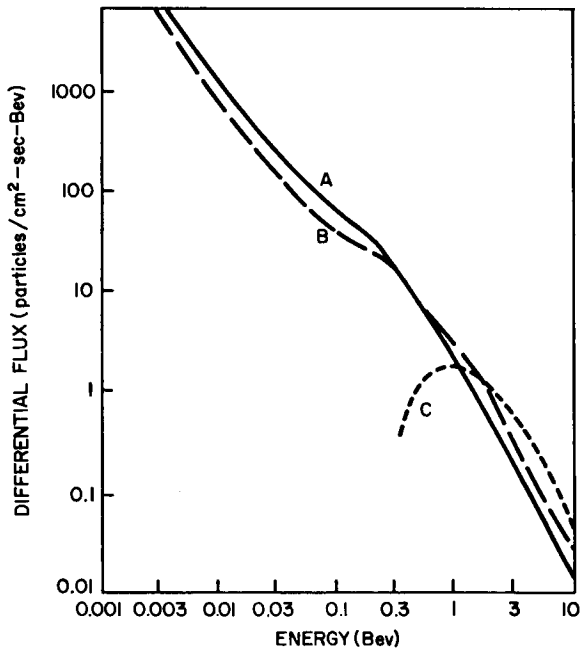


FIGURE 1.—Energy spectrum of nuclear particles in a small iron meteorite; A—100 gm/cm², B—10 gm/cm², C—primary spectrum (after Arnold, reference 18).

nuclides in the Aroos meteorite.[15,16,17] Figure 1, from Arnold's review[18] of cosmic ray interactions in meteorites, shows the derived spectra of Arnold et al.[14] Geiss et al.[19] have shown that the center of an iron meteorite, $\approx$13 cm in radius, will be bombarded by 1.1 proton equivalents $(E \geq 1$ Bev$)/\text{cm}^2\text{-sec}$ or by 2.7×10^{16} protons/cm^2 in 860×10^6 years (the irradiation period of Aroos*)—the equivalent of several weeks irradiation in a terrestrial synchro-cylotron. Detailed studies are presently being made of nuclear production rates at various depths in simulated meteorites† (reference 20 and others).

Geiss et al.[19] and Honda et al.[15] have shown the spectrum and intensity of cosmic radiation to have been constant to within at least a factor of 2 over the last 10^7 years. The variation within the past 10^5 years has apparently been less than 10%,[21] although Voshage[22] has indicated that between 10^6 and 4×10^7 years ago the intensity, more or less suddenly, increased by a factor of 1.3 to 1.8. Over the past 10^8 to 10^9 years the intensity has been similar to that at present.[14]

Whether the sun is a significant source of cosmic radiation is thus far uncertain. Stoenner et al. (reference 23†) found little difference ($\leq$15 percent) in cosmic ray flux near the sun or in the asteroidal belt. Fireman and DeFelice,[24] however, reported that the flux near the earth is about twice that in the asteroidal belt. The significance of the sun as a cosmic ray source should be clarified in the near future as data on nuclear reactions in artificial satellites accumulate (references 25, 26, and 27‡). Fireman, [28] on the basis of material recovered from Discoverer XVII satellite, suggested that solar flares contain an appreciable amount of tritium. Very recently Fireman, DeFelice, and Tilles§ have reported on some cosmic ray and solar proton induced activities in a number of rocket and satellite fragments. Their findings indicate that the November 12, 1960, solar flare contained approximately 1% tritium and that the earth's radiation belts contain about 1 to 4% trapped tritons. Fireman et al. have concluded that the trapped tritium is more energetic than is flare tritium and has a mean lifetime of about 2 months. It would be instructive to examine the D/H ratio in some of these rockets and satellites. The magnitude of any anomalously large ratio (greater than the terrestrial value of 0.00014) in concert with the observed T/H ratio would give some information on hydrogen trapping mechanisms, solar chromospheric thermonuclear reactions, etc.

Questions of some interest are whether the rate of space erosion of meteorites is high and whether any direct evidence of impacts between large meteorites (asteroids) can be obtained. In an extensive review of meteorite ages, Anders[29] concluded that preterrestrial erosion is relatively unimportant in the history of meteorites. Atmospheric ablation, on the other hand, does affect the size and shape[30,31,32,33] of recovered meteorites (figs. 2a and 2b). Figure 2b is an illustration of a case where a meteorite (Carbo) was fractured late in its history—probably on entering the atmosphere.[29] On the basis of exposure ages (see below) Vilcsek and Wänke[34] concluded that the giant Sikhote Alin iron meteorite had undergone several breakups while in space.

Anders[29] has reviewed the subject of meteorite ages in great detail, and only a few of the significant points will be mentioned here. In its most general form, the cosmic ray exposure age τ of a meteorite is given by:

$$\tau = K(\phi R/\phi S)/(S/R), \qquad (4)$$

where K is a proportionality constant and $\phi R/\phi S$ is the relative production ratio of some stable S and radioactive R species. A very comprehensive discussion of types of age determinations is presented in Anders' review.

Thus far, it would appear[29] that many of the iron meteorite age determinations fall into two groups: 2.5 to 3×10^8 and 5 to 6×10^8 years. The ages of the stone meteorites are, in general, much lower than those of the irons, with discrete groups being more apparent. Kirsten et al.[35] presented

*Lipschutz, M. E., "Al^{26} and Ti^{44} in Some Iron Meteorites and Their Exposure Ages," to be presented at the American Geophysical Union meeting in April 1963.

†Davis, R., Jr., et al., "Cosmic Ray Produced Ar^{37} and Ar^{39} Activities in Recently Fallen Meteorites," to be published.

‡Also, see program of meeting on "Scientific Results From a Naturally Recovered Satellite," *Trans. AGU* 43(4): 423, 1962.

§Fireman, E. L., DeFelice, J., and Tilles, D., "Tritium and Radioactive Argon and Xenon in Meteorites and Satellites," IAEA Preprint No. SM 33/44 (1962).

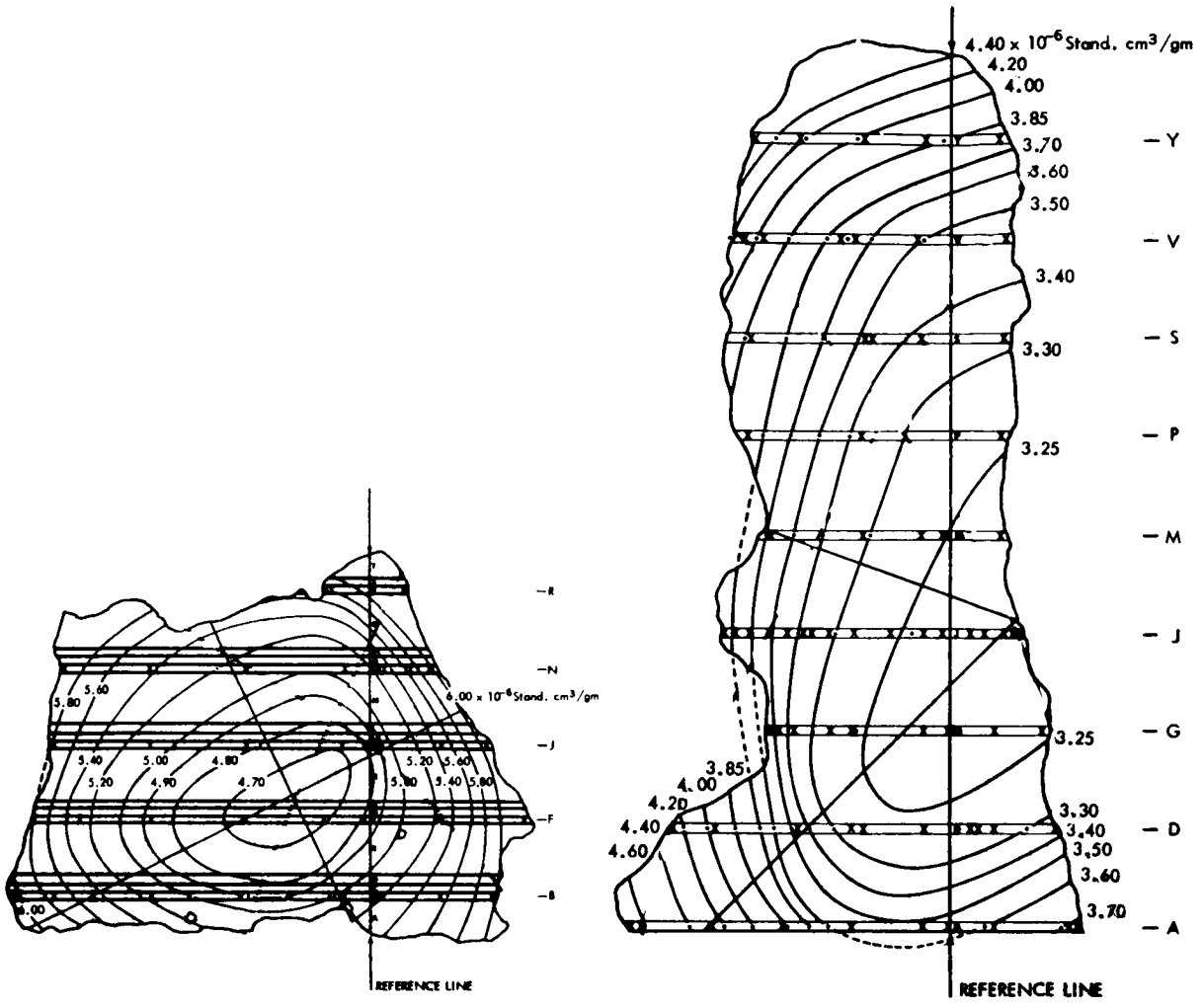


FIGURE 2.—Contours of constant He^3 in two iron meteorites; crosses (×) indicate actual He^3 determinations, and dots (.) indicate interpolated values. (a) Grant iron meteorite (reference 31); (b) Carbo iron meteorite (reference 32); open contour lines to the right indicate a recent fracture.

data that indicate groups at 0 to 0.1×10^8 and 0.15 to 0.3×10^8 years (fig. 3). One stone meteorite has been irradiated by cosmic rays for only about 10^5 years,[36] while the "oldest" meteorite, an iron, apparently has been irradiated for about 15×10^8 years.[17] It is usually assumed that the exposure age represents the time elapsed since the individual meteorites were broken from their parent bodies. Also assumed is that effects due to subsequent space erosion and breakup are insignificant. If, as it now appears, these groups are real, they represent major collisions in the asteroidal belt around 10^7, 2×10^7, 3×10^8, and 5×10^8 years ago. Future effort will undoubtedly reveal whether these groupings are the result of breakups or are due to

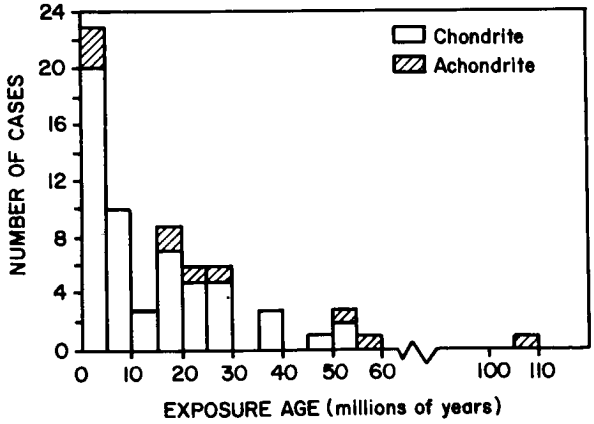


FIGURE 3.—Cosmic ray exposure of some stony meteorites (after Kirsten et al., reference 35).

some other cause (e.g., space erosion). It might be mentioned at this point that some recent investigations* (references 37, 38, and others) have shown that meteorites are quite durable objects. one iron has been found to have been on the earth for $\approx 3 \times 10^5$ years,[15] and two stones for $\geq 20{,}000$ years.[37,38] In most other cases, meteorites' residence on the earth has been longer than 300 to 1500 years. Considering the effects of terrestrial weathering, such "ages" are indeed impressive.

CONCLUDING REMARKS

In this discussion we have reviewed some recent studies of interplanetary radiation and its effects on solid objects. If an attempt is made to extrapolate into the next few years, we can expect to learn a good bit more of the relatively recent ($\leq 10^6$ to 10^8 years) history of our solar system. In particular, the importance of the moon as a source of meteoritic material, and the number of collisions in the asteroidal belt, may be settled. Too, we may begin to estimate the rate of collision of large objects with dust and larger "debris" and to set narrower limits for temporal and spatial cosmic ray variations. By then we will no doubt be studying actual lunar and planetary material in the laboratory. The experience and knowledge gained from previous research on meteoritic extraterrestrial material will be invaluable.

REFERENCES

1. Prior, G. T., "Catalogue of Meteorites," ed. by M. H. Hey, 2nd Ed., London: British Museum, 1953.
2. Heide, F., "Kleine Meteoritenkunde," 2nd Ed., Berlin: Springer, 1957.
3. Krinov, E. L., "Principles of Meteoritics," Transl. by I. Vidziunas, Oxford: Pergamon Press, 1960.
4. Shoemaker, E. M., "Penetration Mechanics of High Velocity Meteorites, Illustrated by Meteor Crater, Aroniza," in: *Rept. 21st Session Norden, Internat. Geolog. Cong., Copenhagen, 1960*, Copenhagen: Det Berlingske Bogtrykkeri, 1960, Part 18, pp. 418–434.
5. Lipschutz, M. E., and Anders, E., "The Record in the Meteorites–IV. Origin of Diamonds in Iron Meteorites," *Geochim. Cosmochim. Acta* **24**(1/2): 83–105, June 1961.
6. Brown, H., "The Density and Mass Distribution of Meteoritic Bodies in the Neighborhood of the Earth's Orbit," *J. Geophys. Res.* **65**(6): 1679–1683, June 1960.
7. Brown, H., "Addendum: The Density and Mass Distribution of Meteoritic Bodies in the Neighborhood of the Earth's Orbit," *J. Geophys. Res.* **66**(4): 1316–1317, April 1961.
8. Goles, G. G., Fish, R. A., and Anders, E., "The Record in the Meteorites–I. The Former Environment of Stone Meteorites as Deduced from K^{40}–Ar^{40} Ages," *Geochim. Cosmochim. Acta* **19**(3): 177–195, 1960.
9. Urey, H. C., "Primary and Secondary Objects," *J. Geophys. Res.* **64**(11): 1721–1737, November 1959.
10. Ceplecha, Z., Rajchl, J., and Sehnal, L., "New Czechoslovak Meteorite 'Luhy'," *Bull. Astronom. Inst. Czechoslovakia* **10**(4): 147–148, July 1, 1959.
11. Houtermans, F. G., and Jager, E., et al., "Messungen der Thermolumineszenz als Mittel zur Untersuchung der Thermischen und der Strahlungsgeschichte von Naturlichen Mineralien und Gesteinen," *Amalen der Physik* **20**(1/6): 283–292, 1957.
12. Heymann, D., and Fluit, J. M., "Sputtering by 20-Kev Ar^+ Ions at Normal Incidence on Meteorites," *J. Geophys. Res.* **67**(7): 2921–2924, July 1962.
13. Goel, P. S., "Calculation of Production Rates of Specific Nuclides in Iron Meteoroids," in: *Researches on Meteorites*, ed. by C. B. Moore, New York: Wiley, 1962, pp. 36–67.
14. Arnold, J. R., Honda, M., and Lal, D., "Record of Cosmic-Ray Intensity in the Meteorites," *J. Geophys. Res.* **66**(10): 3519–3531, October 1961.
15. Honda, M., Shedlovsky, J. P., and Arnold, J. R., "Radioactive Species Produced by Cosmic Rays in Iron Meteorites," *Geochim. Cosmochim. Acta*, **22**(2/4): 133–154, March 1961.
16. Stauffer, H., and Honda, M., "Cosmic-Ray-Produced Stable Isotopes in Iron Meteorites," *J. Geophys. Res.*, **67**(9): 3503–3512, August 1062.
17. Signer, P., and Nier, A. O. C., "The Measurement and Interpretation of Rare Gas Concentrations in Iron Meteorites," in: *Researches on Meteorites*, ed. by C. B. Moore, New York: Wiley, 1962, pp. 7–35.
18. Arnold, J. R., "Nuclear Effects of Cosmic Rays in Meteorites," in: *Annual Review of Nuclear Science*, ed. by E. Segre, G. Friedlander, and W. E. Meyerhof, Palo Alto, Calif.: Annual Reviews, 1961, vol. **11**, pp. 349–370.
19. Geiss, J., Oeschger, H., and Schwarz, U., "The History of Cosmic Radiation as Revealed by Isotopic Changes in the Meteorites and on the Earth," *Space Sci. Rev.*, **1**(2): 197–223, October 1962.
20. Honda, M., "Spallation Products Distributed in a Thick Iron Target Bombarded by 3-Bev Protons," *J. Geophys. Res.*, **67**(12): 4847–4858, November 1962.
21. Hemann, D., and Schaeffer, O. A., "Exposure Ages of Some Iron Meteorites," *J. Geophys. Res.* **66**(8): 2535–2536, August 1961 (abstract).
22. Voshage, H., "Eisenmeteorite als Raumsonden fur die Untersuchung des Intensitatverlaufes der Kosmischen Strahlung Wahrend der Letzten Milliarden Jahre," *Zeitschrift fur Naturforschung* **17**a(5): 422–432, May 1962.

*See asterisked footnote on p. 109. Also, see data in Ref. 29 taken from E. Vilcsek and H. Wänke, whose paper is to be published in *Zeitschrift für Naturforschung*.

23. STOENNER, R. W., SCHAEFFER, O. A., and DAVIS, R., Jr., "Meteorites as Space Probes for Testing the Spatial Constancy of Cosmic Radiation," *J. Geophys. Res.* **65**(10): 3025–3034, October 1960.
24. FIREMAN, E. L., and DEFELICE, J., "Argon 37, Argon 39, and Tritium in Meteorites and the Spatial Constancy of Cosmic Rays," *J. Geophys. Res.* **65**(10): 3035–3041, October 1960.
25. KEITH, J. E., and TURKEVICH, A. L., "Radioactivity Induced in Discoverer 17 by Solar-Flare Protons," *J. Geophys. Res.* **67**(12): 4525–4532, November 1962.
26. SCHAEFFER, O. A., and ZAHRINGER, J., "Solar Flare Helium in Satellite Materials," *Phys. Rev. Letters* **8**(10): 389–390, May 15, 1962.
27. WASSON, J. T., "Radioactive Cobalt and Manganese in Discoverer 17 Stainless Steel," *J. Geophys. Res.* **67**(9): 3513–3517, August 1962.
28. FIREMAN, E. L., "Tritium in Meteorites and in Recovered Satellite Material," in: *Sympos. on the Detection and Use of Tritium in the Physical and Biological Sciences, Vienna*, 1961, Vienna: International Atomic Energy Agency, 1962, Vol. **1**, pp. 69–74.
29. ANDERS, E., "Meteorite Ages," *Rev. Mod. Phys.* **34**(2): 287–325, April 1962.
30. FIREMAN, E. L., "Distribution of Helium-3 in the Carbo Meteorite," *Nature*, **181**(4625): 1725, June 21, 1958.
31. HOFFMAN, J. H., and NIER, A. O., "Production of Helium in Iron Meteorites by the Action of Cosmic Rays," *Phys. Rev.* **112**(6): 2112–2117, December 15, 1958.
32. ———, and ———, "The Cosmogenic He^3 and He^4 Distribution in the Meteorite Carbo," *Geochim. Cosmochim. Acta*, **17**(1/2): 32–36, August 1959.
33. ———, and ———, "Cosmic-Ray-Produced Helium in the Keen Mountain and Casas Grandes Meteorites," *J. Geophys. Res.* **65**(3): 1063–1068, March 1960.
34. VILCSEK, E., and WANKE, H., "Das Strahlungsalter der Eisenmeteorite aus Chlor-36-Messungen," *Zeitschrift fur Naturforschung* **16**a(4): 379–384, April 1961.
35. KIRSTEN, T., KRANKOWSKY, D., and ZAHRINGER, J., "Edelgas und Kaliumbestimmungen einer Grosseren Zahl von Steinmeteoriten," *Geochim. Cosmochim. Acta* (in press).
36. ANDERS, E., "Two Meteorites of Unusually Short Cosmic-Ray Exposure Age," *Science* **138**(3538): 431–433, October 19, 1962.
37. GOEL, P. S., and KOHMAN, T. P., "Cosmogenic Carbon-14 in Meteorites and Terrestrial Ages of 'Finds' and Craters," *Science*, **136**(3519): 875–876, June 8, 1962.
38. SUESS, H. E., and WANKE, H., "Radiocarbon Content and Terrestrial Age of Twelve Stony Meteorites and One Iron Meteorite," *Geochim. Cosmochim. Acta*, **26**, 475–480, April 1962.

ALUMINUM-26 IN IRON METEORITES AND THEIR COSMIC-RAY EXPOSURE AGES

MICHAEL E. LIPSCHUTZ*

AND

EDWARD ANDERS

Enrico Fermi Institute for Nuclear Studies, University of Chicago

The Al^{26} and Ti^{44} contents of 10 iron meteorites were determined by γ–γ coincidence counting of separated fractions. The Al^{26} contents, ranging from 0 to 2.84 dpm/kg, were lower by factors of 1.5 to 3 than literature values obtained by β-counting. This discrepancy is apparently caused by a β-emitting impurity that interferes in β-counting, but not in the γ–γ coincidence method. From the Ne^{21} content of these meteorites, the following cosmic-ray exposure ages, in million years, were calculated (relative to Grant = 640 m.y.): Arispe = 110, Aroos = 950, Duchesne = 170, Norfolk = 530, Odessa = 360, Sandia Mountains $\geq$ 70, Santa Apolonia = 670, Sikhote-Alin = 330, Treysa = 400. In most cases, these values agree rather well with previously determined K^{40}/K^{41} and Cl^{36}/Ar^{36} ages.

Except for the observed falls Aroos, Sikhote-Alin, and Treysa, none of the above meteorites contained any detectable Ti^{44} activity. Evidently the terrestrial ages of these meteorites are long compared to the half-life of Ti^{44}, ~150 y. The absence of Ti^{44} activity in Norfolk shows that it cannot have been associated with a 1907 fireball, as had been previously supposed.

I. INTRODUCTION

During the past several years cosmic ray exposure ages of iron meteorites have been determined by many investigators. However, the ages obtained for the same meteorite by different methods are often in disagreement—differences of a factor of two being quite common with occasional differences as high as a factor of ten (Anders, 1962). These disagreements may be due to a number of causes which include: diffusion losses of gaseous nuclides (Fireman and De Felice, 1960, Geiss, Hirt, and Oeschger, 1960, Bainbridge, Suess, and Wänke 1962); terrestrial ages comparable to the half-lives of the radionuclide measured (Fisher and Schaeffer 1960, Honda, Shedlovsky, and Arnold 1961, Vilcsek and Wänke, 1961); variations in nuclide production rates due to shielding (Hoffman and Nier, 1958, 1959, 1960); actual age differences in various portions of the meteoroid owing to the exposure of fresh surfaces by multiple collisions (Vilcsek and Wänke, 1961, 1963); and secondary production processes which depend on variables other than exposure time [e.g. thermal neutron capture by lawrencite (Goel, 1962)—$FeCl_2$].

In order to develop a dating method which would be relatively insensitive to at least some of these effects we isolated Al^{26} and, in addition, several other cosmogenic radionuclides, including Ti^{44}, from iron meteorites. The Al^{26} contents, combined with Ne^{21} measurements, were used to calculate cosmic ray exposure ages.

Aluminum-26 ($t_{1/2} = 7.4 \times 10^5$ years; Rightmire, Kohman, and Hintenberger, 1958) decays primarily (84%) by positron emission (Kohman, Rightmire, Ehmann, and Simanton, 1957) while Ti^{44} decays by electron capture to 4 hr-Sc^{44} which, in turn, decays largely (93%) by positron emission (Blue and Bleuler, 1955). Thus both of these nuclides can be counted by γ–γ coincidence spectrometry. This technique discriminates very powerfully against β-emitters and most γ-emitters, so that the samples need not be processed to the same high degree of radiochemical purity required

*On active duty with U.S. Army.

in counting. In fact, this technique has been utilized by several investigators (Anders 1960; Gfeller, Houtermans, Oeschger and Schwarz, 1961; Mayer-Böricke, Biswas, and Gentner, 1962; Fireman, Fazio, and DeFelice, 1963) for the quantitative determination of Al^{26} by non-destructive counting of various stony meteorites.

Since Al^{26} and Ne^{21} are 30–35 mass units removed from the principal target nucleus, Fe^{56}, they are produced mainly by high-energy cosmic-ray primaries. Therefore, their production cross sections should have nearly the same energy-and depth-dependence. It is true that both nuclides may be produced in high yield by spallation of phosphorus or sulfur. However, the phosphorus content of iron meteorites is usually less than 0.5% and thus its contribution to the meteoritic contents of Al^{26} and Ne^{21} should be negligible. Sulfur occurs mainly in the form of troilite (FeS) nodules which are normally avoided in sampling.

II. EXPERIMENTAL

1. Counting

The application of γ–γ coincidence spectrometry to the measurement of β^+ emitters has been described by Anders (1960) and by Viste and Anders (1962). Briefly, the technique involves coincidence counting of the 0.51 Mev photons produced in 180° angular correlation by the annihilation of positrons emitted by the radionuclide.

Two coincidence spectrometers were used in this study. The first consisted of two 3.8×3.8 cm. NaI (T1) scintillation crystals mounted on EMI type 9536 photomultiplier tubes.* The resolution of these detectors for the 662 Kev gamma ray of Cs^{137} were 10.1 and 9.5% and their counting efficiencies were 8.3 and 8.1%. The positron counting efficiency of the coincidence mode was found to be 1.49±0.04% as determined by calibration with a National Bureau of Standards Na^{22} sample. This standard, NBS No. 4922, had a positron specific activity of $7.81\pm0.16\times10^6$ β^+/min./ml. in October 1954. In making decay corrections a Na^{22} half-life of 2.58 years was used.

*The 90 gram Toluca troilite nodule was counted by 6.4×3.8 cm copper-canned NaI (Tl) crystals spaced 2.3 cm apart. The counting yield was 1.75±0.06%.

The second spectrometer system is shown schematically in figure 1. Detectors 1 and 2 (D1 and D2) were two 3.8×5.1 cm NaI (T1) scintillation crystals on Dumont type 6292 photomultiplier tubes. Their resolutions were 8.4 and 8.7% and their counting efficiencies, 13.0 and 12.7% respectively. The coincidence counting efficiency was 2.43±0.07%.

Samples were counted in cylindrical lucite capsules, 20.5 mm o.d., 14.0 mm i.d. and 1.7–4.2 mm. thick. They were covered at either end with 0.25 mm electrolytic-copper discs sealed on with Eastman 910 adhesive. A constant spacing of 6.30 mm between the detectors was maintained by means of a lucite sample holder.

The detectors were surrounded by an annular anticoincidence Geiger counter inside of a massive shield constructed from concentric sleeves of cast iron, electrolytic copper and high-purity bismuth. Between the bismuth and the anticoincidence counter were sleeves of 0.225 cm thick high-purity cadmium and 0.025 cm thick electrolytic copper, to degrade bismuth X-rays. The doors, hinged at either end, were made of cast iron and bismuth slabs.

The signals from the detectors were fed into a dual amplifier (Nuclear Data ND500), set so as to pass gamma rays in the 475–545 Kev energy

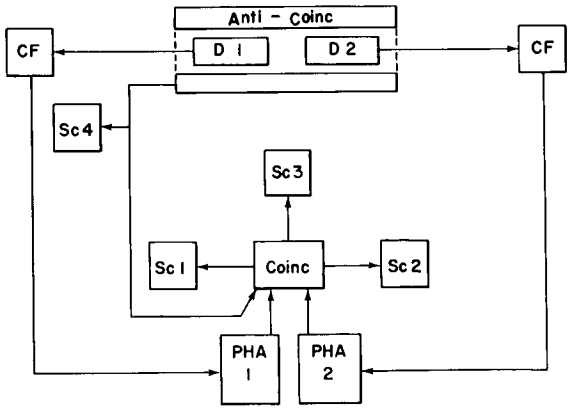


FIGURE 1.—Block diagram of counting equipment. Pulses from the detectors (D1 and D2), after passing through the cathode followers (CF), enter the pulse height analyzers (PHA). Signals from these are fed into the coincidence circuit (Coinc) where they may be blocked by anti-coincidence-ring pulses (Anti-coinc) or establish a mutual coincidence in the energy region 475–545 Kev. The scalers (Sc) record pulses from the individual units.

TABLE I.—*Description of Meteorite Samples Investigated*

Meteorite and year of fall	Recovered mass (kg)	Source	Remarks
Arispe	398	Mr. G. I. Huss	Saw cuttings.
Aroos (1959)	150	Dr. M. Honda	Previously separated Al_2O_3 and TiO_2 (Honda and Arnold, 1961).
Canyon Diablo	>30,000	Dr. T. P. Kohman	Previously separated TiO_2.
Duchesne	22.8	Mr. G. I. Huss	Saw cuttings.
Grant	480	Dr. M. Honda	Previously separated Al_2O_3 (Honda, et al, 1961).
Santa Apolonia	1315	Mr. G. I. Huss	Saw cuttings.
Norfolk (1907?)	23	Dr. B. Mason	Saw cuttings.
Odessa	>10,000	Dr. M. Honda	Previously separated Al_2O_3 (Honda et al, 1961).
Sandia Mountains	45.4	Mr. G. I. Huss	Saw cuttings.
Sikhote Alin (1947)	>23,000	Dr. D. Heymann	Meteorite solution (Heymann and Schaeffer, 1961).
Sikhote Alin III (1947)	>23,000	Dr. M. Honda	Previously separated TiO_2.
Toluca (Troilite)	>20,000	Mr. G. I. Huss	Single nodule.
Treysa	63	Dr. O. A. Schaffer	Shavings.
Treysa II (1916)	63	Dr. M. Honda	Previously separated TiO_2.

interval. The anticoindence signal was used to gate the output of one of the amplifiers before it entered the coincidence circuit.

Total sample counting times ranged from 12,000 to 40,000 minutes. Individual samples were counted alternately with background for periods of 1000–3000 minutes. The dual amplifier was usually calibrated daily and was seldom found to drift as much as 1%. On those occasions when drift was greater than 1%, the count was discarded.

Because of the constant spacing between crystals, background count rates were nearly constant. Normally, the gated detector registered 1.80 ± 0.05 c/m and the ungated detector, 2.48 ± 0.05 c/m. On those occasions when one of the detectors varied beyond these limits, the count was discarded. The coincidence background was $3.78\pm0.15\times10^{-3}$ c/m.

2. Chemical Procedure

Table I lists sources and descriptions of the individual samples. Iron meteorite samples of 60–190 grams (usually in the form of saw cuttings) were dissolved under reflux in concentrated hydrochloric and nitric acids, or in 2M sulfuric acid if Cl^{36} was to be determined. In the latter case 12.9 mg of Cl^- carrier and 0.40 mg of Cl^{37} spike were added before solution of the sample was begun. Magnetic particles still present in the residue when solution was apparently complete were dissolved in HNO_3 and the solution from this step was added to the main solution. Those samples which had been dissolved in sulfuric acid were then oxidized under reflux with 15% H_2O_2 and the chlorine was precipitated as AgCl and filtered.

After removal of about $\frac{1}{10}$ of the filtrate for later analysis, the following carriers were added: Ti (50.5 mg), Mn (13.1 mg), Al (300 mg), and Be (9.4 mg). No carrier was added for Ni as a large amount of nickel is present in all iron meteorites.

Owing to anodic oxidation of Fe^{++} to Fe^{+++} it proved impossible to remove iron by electrolysis. For this reason, most of the Fe^{+++} was removed by solvent extraction with isopropyl ether from 8N HCl solution (Sandell, 1944, pp. 28–29).

The solution, at this point usually about 8 liters, was concentrated by distillation of hydrochloric acid, and the residue was dissolved in distilled water and diluted to 1–2 liters until the

TABLE II.—*Results of Al^{26} Determinations*

Sample	Weight of sample (g)	Chemical yield (%)	Net count rate (c/1000m)	Disintegration rate† (dpm/kg)
Arispe*	130.2	79	2.70±0.56	1.64±0.36
			‡2.16±0.73	
Duchesne*	190.1	89	6.32±0.70	2.02±0.22
Norfolk	80.0	65	3.03±0.64	2.84±0.60
Sandia Mountains*	106.6	96	−0.60±0.43	−0.31±0.22
Sikhote Alin I	55.8	85	0.77±0.38	0.80±0.39
Santa Apolonia*	100.2	97	1.17±0.57	0.66±0.32
Toluca Troilite	90.1		1.7 ±2.1	1.5 ±1.8
Treysa I*	100.0	95	4.67±0.61	2.66±0.35
Aroos**	254	90	6.78±0.70	1.51±0.15
Grant**	459	54	9.62±0.84	1.90±0.17
Odessa**	3540	33	12.62±1.16	0.56±0.048

*After the sample was dissolved, 10% of the solution was reserved for later analysis.
**These samples were previously chemically separated by Honda and coworkers at La Jolla.
†84% β^+ emission.
‡1.49%-counting efficiency.

solution was about 1N in H_2SO_4. This solution was electrolyzed with a mercury cathode* at 10–20 and a voltage of 3.4 volts until loss of color indicated that removal of Ni^{++} and Co^{++} was complete.

The solution from the electrolysis was then concentrated to 1.8N in sulfuric acid and titanium was precipitated by 4% *p*-hydroxyphenylarsonic acid solution. This procedure is quantitative for amounts of Ti as small as 17 mg (Simpson and Chandlee, 1938). The precipitate was ignited to TiO_2 well-ventilated hood and fumed to dryness with HF solution to volatilize silicon as SiF_4. After another ignition the chemical yield of the titanium was determined gravimetrically and the TiO_2 was then placed in a sample mount and counted. No correction for meteoritic Ti was necessary as Wänke (1960) has found this to range from 0.6 to less than 5 ppm in 4 iron meteorites.

The filtrate, after removal of titanium, was evaporated to near dryness and poured into concentrated NaOH solution to precipitate insoluble hydroxides (including U and Th). After centrifuging the solution and washing the precipitate with 20% NaOH, the supernate and washes were neutralized with concentrated HCl to a pH of 8.0–8.6 and centrifuged. The amphoteric hydroxide precipitate was then washed with a buffered wash and dissolved in concentrated HCl. Aluminum was then precipitated with 8-hydroxyquinoline (Flagg, 1948) and ignited to Al_2O_3. After gravimetric determination of the chemical yield the Al_2O_3 was compacted on a hydraulic press into small pellets using mineral oil as a binder, and counted. The yield was not corrected for aluminum originally present since its content in iron meteorites is very low.

III. RESULTS

The data obtained are listed in tables II and III. The uncertainty in each case is the statistical standard deviation of counting since all other experimental uncertainties are trivial compared with this value.*

*See Maxwell and Graham (1950) and Page, Maxwell and Graham (1962) for reviews of the method.

*The absolute Sikhote Alin I Al^{26} disintegration rate may be somewhat uncertain. This sample was received as a solution (table I) to which unknown amounts of Al and Ti carriers had been added. The aluminum chemical yield was arbitrarily estimated at about 85%.

This estimation should not be in error by more than about 10%. The Ti yield was estimated by averaging all of the other Ti chemical yields (except that of Sandia Mountains). This estimate, too, should not be in error by more than 10%.

TABLE III.—*Results of Ti^{44} Measurements*

Sample	Weight of sample (g)	Chemical yield (%)	Net count rate (c/0000m)	Disintegration rate† (dpm/kg)
Arispe*	130.2	88	‡0.12±0.64 0.15±0.64	0.076±0.26
Norfolk	80.0	92	0.34±0.75	0.20 ±0.44
Santa Apolonia*	100.2	89	0.15±0.52	0.083±0.29
Sikhote Alin I	55.8	91	1.27±0.53	1.11 ±0.46
Treysa I*	100.0	100	‡1.52±0.73	1.23 ±0.59
Aroos**	254	90	6.9 ±1.1	2.28 ±0.36
Treysa II**				2.32 ±0.43
Canyon Diablo***	219	92		
Sikhote Alin III**			2.94±0.87	1.46 ±0.43

*After the sample was dissolved, 10% of the solution was reserved for layer analysis.
**These samples were previously chemically separated by Honda and coworkers at La Jolla.
***This sample was previously chemically separated by J. Kaye at the Carnegie Institute of Technology.
†93% positron emission.
‡1.49% counting efficiency.

TABLE IV.—*Comparison of Al^{26} and Ti^{44} Samples by β and γ–γ Counting*

Sample	γ–γ counting (dpm/kg)	β counting (Honda et al.)* (dpm/kg)	β/γ–γ	β counting (other)* (dpm/kg)
Al^{26}				
Aroos	1.51±0.15	[1]3.6 ±0.4	2.4±0.3	
Grant	1.88±0.16	[2]3.67±0.24	2.0±0.3	[3]4.70±0.70
Odessa	0.56±0.048	[2]1.17±0.06	2.1±0.2	[4]5.5 ±1.0
				[6]0.62±0.09
				[6]1.39±0.12
Sikhote Alin	0.80±0.39			[5]3.1 ±0.7
				[6]1.73±0.18
				[7]1.74±0.9
				[7]0.99±0.18
Treysa	2.66±0.35			[5]7.4 ±1.2
Ti^{44}				
Aroos	2.28±0.36	[1]4.4 ±0.4	1.9±0.3	[7]3.18±0.46
Sikhote Alin III	1.46±0.43	[8]1.8 ±0.4	1.2±0.4	
Treysa II	2.32±0.43	[8]3.0 ±0.7	1.3±0.4	

*Corrected value (see text).
[1] Honda and Arnold, 1961.
[2] Honda, Shedlovsky and Arnold, 1961.
[3] Crevecoeur and Schaeffer, 1962.
[4] Ehmann and Kohman, 1958.
[5] Sammet and Herr, 1963.
[6] Chakrabartty, 1961.
[7] Kaye, 1963.
[8] Honda, 1962.

1. Comparison of Samples by Both Beta and Coincidence Counting

Unfortunately, our results disagree with all previous determination of Al^{26} in iron meteorites, and these determinations, in turn, disagree among themselves (table IV). All, except our results, were obtained by beta counting.* Before we can calculate any cosmic-ray ages from our Al^{26} data we must establish which os these values most nearly represent the true Al^{26} content of meteorites.

The disagreement is most crucial between our results and those of Honda et al., since both measurements are made on identical samples (supplied through the courtesy of Dr. M. Honda). Evidently, Honda's results, obtained by β-counting, are systematically higher than our γ–γ measurements. This discrepancy might be explained in two ways.

(a) Errors in Absolute Calibration of the γ–γ or β counters.

Honda et al. (1961) calibrated Geiger counters used in their work by National Bureau of Standards beta emitters and natural K^{40} samples, with appropriate corrections for self-absorption. Observed counting rates were converted to absolute disintegration rates by applying corrections for sample length, self-absorption, chemical yield, and decay modes.

Our γ–γ spectrometers were calibrated by a National Bureau of Standards Na^{22} source. Since Na^{22} is a positron emitter it provided a simple and direct means of establishing the counting efficiency of our spectrometers for other positron emitters such as Al^{26} and Ti^{44}-Sc^{44}.

That this error is not the source of the discrepancy is indicated by the inequality of the ratio of the β to γ–γ count rates (table IV, column 4). This ratio varies over a considerable range from 1.2 to 2.4. An error in calibration would result in constant ratios.

(b) Impurities.

Beta-emitting impurities in the samples of Honda et al. would increase the β count rate relative to the γ–γ rate. Unfortunately, owing to the low counting rates it was not feasible to look for such impurities by taking beta absorption curves. Three years after the original measurement of the Al^{26} beta activities, Honda recounted his Odessa and Aroos Al_2O_3 samples and found that their specific activities had decreased by about 20–30% (private communication) due to decay of some contamination. Further recycling by Honda subsequent to our measurements actually *increased* the specific activities—showing that some contaminant was being introduced.

The γ–γ coincidence method, on the other had, is much less sensitive to contamination since only positron emitters, which are rare, and nuclides with two coincident gamma rays of more than 0.5 Mev energy could interfere. While we have no conclusive evidence that our samples were entirely free of such contaminants, we have reason to believe that they were not present in significant amounts. Four of our samples (tables II and III) gave activities essentially equal to zero; and if the average activity of these samples, -0.08 ± 0.28 c/1000 minutes is taken to represent our contamination level, the correction to our results

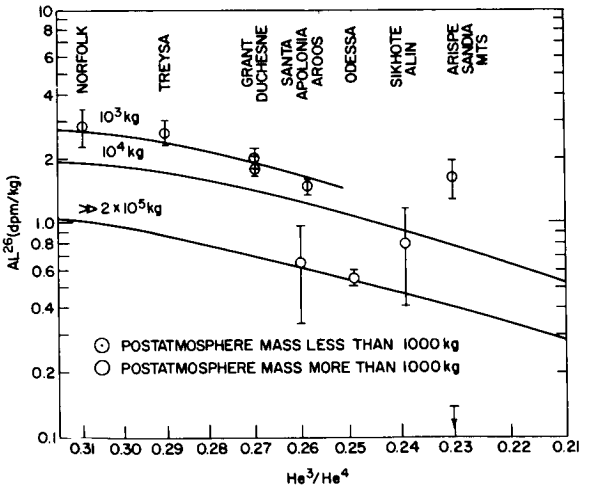


FIGURE 2.—Al^{26} content of iron meteorites as a function of depth and size. The curves are derived from Signer and Nier's (1959) data for Ne^{21}. With the exception of Arispe, the meteorites lie in the region of the diagram appropriate to their mass.

*The original values published for Grant and Odessa were high by a factor of 1.17 owing to a numerical error (Honda et al., 1961). The values in table IV have been lowered by this factor which should also be applied to the published result for Williamstown.

TABLE V.—*Cosmogenic Noble Gas Contents*

Meteorite	Ne^{21} 10^{-8} cc (STP)g	He^3/Ne^{21}	He^3/He^4	He^4/Ar^{38}	Reference
Arispe	1.0	92	0.23	77	(2)
Aroos	8.15	80.5	0.259	58.5	(3)
Duchesne	2.0	65	0.27	53	(2)
Grant	[1]7.0	[1]76	[1]0.27	[1]59	(3)
Norfolk	8.7	70	0.31	57	(2)
Odessa	1.16	89.0	0.249	61.5	(3)
Sandia Mountains	0.5	100	0.23	100	(2)
Santa Apolonia	2.3	83	0.26	67	(2)
Sikhote Alin	1.44	86.0	0.238	70.0	(4)
Treysa	6.2	80	0.29	63	(2)

[1] Mean value (see text).
[2] Signer, 1963.
[3] Signer and Nier, 1962.
[4] Schaeffer, 1963.

would amount to 5–15%. Further supporting evidence for the validity of our results comes from figure II to be discussed below.

2. Interpretation of the Al^{26} Measurements

There is yet another way to check the reasonableness of the data. In a "normal" meteorite (cosmic-ray age longer than several half-lives of Al^{26}; no decay of Al^{26} since fall), the Al^{26} content is mainly a function of depth. Hence it must be consistent with other depth-indicators, such as the He^3/He^4 or the He^3/He^4 ratios.* Figure II shows a plot of our Al^{26} values against the He^3/He^4 ratios. Actually, these two depth indicators differ in one important respect. The Al^{26} content depends on the *absolute* flux of high-energy particles. At a given depth, it is therefore higher in a small meteorite than in a large one, since the former receives cosmic rays over a solid angle close to 4π, while the solid angle for the latter is considerably smaller, approaching 2π for meteorites of very large size. The He^3/He^4 ratio, on the other hand depends on the *ratio* of the two fluxes (high-energy to total). Consequently, it depends primarily on depth, and only to a lesser degree on size.

To illustrate this effect, we have included in figure II three curves showing the relation between Al^{26} content and He^3/He^4 for spherical meteorites having pre-atmospheric masses of 10^3, 10^4, and $\gg 2\times10^5$ kg. These curves were derived from the data of Signer and Nier (1959) by plotting the relative Ne^{21} production rate versus the He^3/He^4 ratios corresponding to these masses and assuming the Al^{26} production rate to equal that of Ne^{21}. The relative curves were then converted to an absolute scale by normalizing the curve for 2×10^3 kg to the observed Al^{26} and He^3/He^4 values for Aroos and Grant.

The agreement of the Al^{26} values with the curves is very gratifying. All the small meteorites except Arispe fall near the 100 kg curve, and all the large ones, between the 10^4 and 2×10^5 kg curves. The point for Arispe falls in a "forbidden" region. Small meteorites cannot have so low a He^3/He^4 ratio, and large meteorites cannot have so high an Al^{26} content.

We cannot rule out the possibility that the Arispe "anomaly" is merely due to an error in the Al^{26} or He^3/He^4 measurement. The latter is backed up by the He/Ne ratios reported by Signer (1962) and by Voshage and Hintenberger (1963): 163 and 159. Both these values suggest

*We are greatly indebted to Professor Peter Signer for permission to use some of his unpublished data. These preliminary results (table V) have been assigned a conservative uncertainty of ±10% of the absolute isotopic concentration. The He^3/He^4 ratio of these samples is more accurately known, however, and has been used in the discussion which follows.

that Arispe was a very large meteorite, thus supporting the He^3/He^4 result. The Al^{26} result, on the other hand, stands alone, and we have no conclusive evidence that the Al^{26} value is not too high because of contamination. A more pleasing alternative is that Arispe, like Odessa and Sikhote Alin (Vilcsek and Wänke, 1961, 1963), has undergone repeated breakups during its cosmic-ray exposure history. Pieces that were located below the surface until a recent breakup will contain stable noble gases in proportions characteristic of heavy shielding, while its radionuclides, having been produced mainly during the last few half-lives, will occur at the high levels appropriate to surface regions. This point is discussed further in sec. D-4.

The low Al^{26} value for Sandia Mountains may indicate a high terestrial age, or very heavy shielding. The latter explanation is to be preferred according to Occam's principle, since the He^3/He^4 ration independently suggests a large preatmospheric mass.

One comment should be made concerning the troilite nodule that we examined. The 2σ upper limit of 5 dpm Al^{26}/kg of troilite is an order of magnitude lower than expected for surface material. This implies that the nodule came from a depth of at least 50 cm below the preatmospheric surface of the meteroid. Should this particular nodule contain the radiogenic lead reported by Marshall and Hess (1961), Urey and Murthy (1963) and Murthy (1963), then their hypothetical preterrestrial alteration processes must have injected it to at least this depth. However, the preterrestrial origin of this lead has been questioned (Anders, 1963).

It is evident that at least some of the Al^{26} measurements by β-counting gave erroneously high values due to contamination. We cannot prove conclusively that *all* such measurements are likewise in error. However, the values reported in the literature (table IV) are often higher by factors of 1.5 to 3 than our highest values (Norfolk, 2.84 dpm/kg; Treysa, 2.66 dpm/kg). By every available criterion, these two meteorites are small and relatively unshielded. The Cl^{36} content of Treysa (20 dpm/kg) is nearly as high as the highest value ever observed: 25.4 dpm/kg for the Charlotte meteorite (Vilcsek and Wänke, 1963). Consequently, it is highly unlikely that any iron meteorite has a true Al^{26} content appreciably above 3 dpm/kg. For large iron meteorites, the limit should be even lower. Most of the published Al^{26} values exceed these limits, and we therefore believe that these results should be confirmed by coincidence counting.

3. Ti^{44} Measurements

Our results are listed in table III. In most of the meteorites examined, the Ti^{44} activity was zero within experimental error. Only the observed falls, Aroos, Sikhote Alin, and Treysa, showed measureable Ti^{44} activity.

TABLE VI.—*Terrestrial Ages*

Meteorite	Te^{44}/Al^{26} 2σ limit	Terrestrial age (years) 2σ limit
Arispe	≤ 0.37	≥ 300
Norfolk	≤ 0.38	≥ 300
Santa Apolonia	≤ 1.0	≥ 90

From the Ti^{44}/Al^{26} ratios in our samples, lower limits can be placed on their terrestrial ages (table VI). We have assumed an initial Ti^{44}/Al^{26} ratio of 1.5—that in Aroos—and a Ti^{44} half-life of 50 years (Honda and Lal, 1961).

From table VI it may be seen that the terrestrial age of Norfolk (supposed to have fallen in 1907—Prior and Hey, 1953) is greater than 300 years. Nininger (1937) has previously pointed out that this meteorite appears somewhat weathered and therefore fell prior to 1907. Schaeffer (personal communication) has observed no 325-year Ar^{31} in this meteorite. Thus it appears most unlikely that Norfolk fell some 56 years ago.

4. Cosmic-Ray Exposure Ages

For an iron meteorite whose Al^{26} and Ne^{21} contents are known, a cosmic-ray exposure age, T, can be calculated:

$$T = k \times Ne^{21}/A^0 \times P(Al^{26})/P(Ne^{21}) \qquad (1)$$

where k is a numerical constant equal to 5.11×10^{16} cm^{-3}, Ne^{21} is the concentration of Ne^{21} in

TABLE VII.—*Cosmic Ray Exposure Ages*

Meteorite	$Al^{26}/Ne^{21}\times10^6$ years	$K^{40}/K^{41*}\times10^6$ years	$Cl^{36}/Ar^{36}\times10^6$ years	$Ar^{39}/Ar^{38**}\times10^6$ years
Arispe	110±25	885±90	†440±30	
Aroos	950±130	920±45	**690±70	650±70
			†540±30	
Duchesne	170±30			
Grant	640±60	695±	**640±70	
Norfolk	530±130	585±	**570±60	
Odessa	360±45	----------	†460±50	
Sandia Mountains	70			
Santa Apolonia	670±340			
Sikhote Alin I	330±110	----------	**310±30	340±30
Treysa I	400±75	610±55	**430±40	370±40
			†450±30	

*Voshage and Hintenberger, 1962.
**Heymann and Schaeffer, 1963.
†Vilcsek and Wänke, 1963.

units of 10^{-1} cm (STP)/g, A^0 is the saturation activity of Al^{26} in units of dpm/kg, while $P(Al^{26})$ and $P(Ne^{21})$ are the production cross sections for Al^{26} and Ne^{21} in an iron meteorite.

This relation implicitly assumes a constant cosmic ray flux, no space erosion, and a depth-independent production ratio for the two nuclides. To a first approximation at least, this last assumption is fulfilled by nuclides similar in mass but quite far removed from the target nucleus, Fe^{56}.

Inasmuch as the relative production ratio of Al^{26} and Ne^{21} has not been directly measured, one must obtain it by indirect methods. The theoretical ratio, at a depth of 100 g/cm², is 0.41 (Arnold, Honda and Lal, 1961). An empirical value can be calculated from the Al^{26} and Ne^{21} contents of a meteorite whose exposure age has been well established by several independent methods.

Of the meteorites investigated, we selected Honda's Grant sample as the normalization standard; partly because the statistical counting error was smaller for this sample than for any of the others, and partly because the exposure age of Grant has apparently been very well established at about 0.6–0.7 AE (Signer and Nier (1962), Heyman and Schaeffer (1963), and Voshage and Hintenberger (1963)). The Al^{26}/Ne^{21} exposure ages listed in table VII are normalized to an assumed age of 0.64 AE for Grant. Any error in the normalization standard will, of course, change the absolute values of the Al^{26}/Ne^{21} ages in table VII while leaving their relative standing unaffected.

Although the cosmogenic gas content of Honda's sample of Grant was not measured directly, it may be quite reliably estimated. The sample, which consisted of saw cuttings from a central cut through the meteorite is probably representative of the slice studied by Hoffman and Nier (1958). From the published Ne^{21} con-

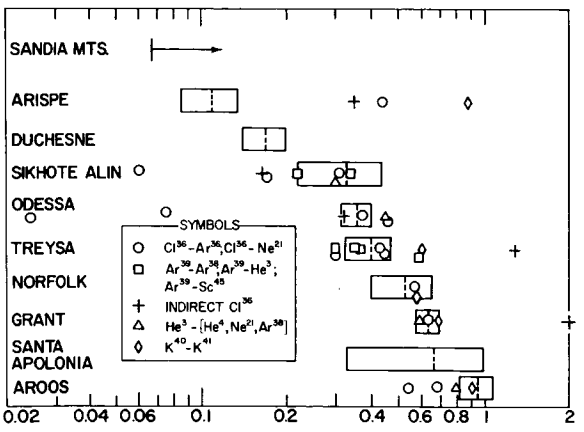


FIGURE 3.—Comparison of Al^{26}/Ne^{21} ages with determinations by other methods. The lengths of the blocks indicate error limits.

tours, an average Ne^{21} content of 7.0 cm^3 (STP)/g is found by numerical integration. Combined with the Al^{26} content from table II, this gives a production ratio, $P(Al^{26})/P(Ne^{21})$ of 0.34.

The Al^{26}/Ne^{21} ages are shown in figure III, together with exposure ages previously deterdetermined by other methods (see the review by Anders (1963) for the data and individual references).

Regrettably, the data do not provide an answer to a question of fundamental importance: Whether meteorites are produced continuously, or in a few discrete events. Some slight tendency toward clustering is evident in the figure, (fig. III) but the sample is too small, and the errors are too large to permit any definite conclusions to be drawn. Nevertheless, a few interesting trends emerge from the data.

In 3 out of 5 cases the Al^{26}/Ne^{21} ages agree quite well with the K^{40}/K^{41} ages. The disagreement for Treysa is moderate, allowing for errors in both the K^{40}/K^{41} method and the normalization. For Arispe, the discrepancy is large, possibly due to multiple breakups of this meteorite, as discussed below.

A comparison of Aroos and Grant, two well-studied meteorites, is instructive. Three methods (K^{40}/K^{41}, He^3/He^4, and Al^{26}/Ne^{21}) consistently show Aroos to be older by a factor of about 1.4, whereas the Cl^{36}/Ar^{36} or Cl^{36}/Ne^{21} methods seem to indicate approximately equal ages for both meteorites. Prior to the availability of our Al^{26}/Ne^{21} data, this discrepancy was blamed on space erosion (Goel and Kohman, 1963). In their point of view, meteoritic material is constantly being eroded by dust. Thus, a short lived nuclide, such as Cl^{36} ($t_{1/2}$=308,000 y) records the cosmic-ray flux during the last million years or so, when most of the overlying material had already eroded away. $K^{40,41}$ and $He^{3,4}$, on the other hand, were produced throughout the entire cosmic-ray exposure era, and were mainly produced at a lower rate, while the meteorite was still heavily shielded. In this event, the Al^{26} age of Aroos should agree with the Cl^{36} ages, since both nuclides have half-lives of less than 10^6 years. They do not. Evidently, the short Cl^{36} ages of Aroos cannot be blamed on space erosion. It seems that some other unknown factor, either chemical or nuclear, raised the Cl^{36} content of Aroos above its proper level. Perhaps the same factor, whatever its nature, was also responsible for the curious depth variation of Cl^{36} in Canyon Diablo, observed by Goel (1963). The inferences drawn from these observations, i.e. strong time variation of cosmic radiation, and rapid space erosion, are rather unorthodox. Perhaps this depth variation should be verified by Al^{26} measurements, before these conclusions are accepted.

The ages for Sikhote, Alin, and Odessa agree rather well, except for the low values. The latter were obtained on different samples of these meteorites, and have been attributed to multiple breakups, which exposed fresh surfaces from time to time. The short Al^{26}/Ne^{21} age of Arispe may well be due to the same cause. Both the He^3/He^4 and Ne^3/Ne^{21} ratio suggest a very large preatmospheric mass for this meteorite. Yet its postatmospheric mass is low (~200 kg); it is not associated with a crater, and its high Al^{26} content (fig. II) also suggests that it spent the final phase of its history as a small, poorly shielded object.

ACKNOWLEDGMENTS

We should like to thank Mr. John Barker for his aid during the latter part of this study. This research was supported in part by the U.S. National Science Foundation and the U.S. Atomic Energy Commission. One of us (M.E.L.) is indebted to the U.S. Rubber Co. for a predoctoral fellowship during a portion of this work.

REFERENCES

ANDERS, E., The Record in the Meteorites II Al^{26} in meteorites and tektites, *Geochim. et Cosmochim. Acta*, **19**, 53–62 (1960).

———, Meteorite Ages, *Revs. Mod. Phys.*, **34**, 287–325 (1962).

———, Meteorite Ages, in *The Solar System, Vol. IV, The Moon, Meteorites and Comets* edited by B. Middlehurst and G. P. Kuiper, University of Chicago Press, pp. 402–495 (1963).

ARNOLD, J. R., HONDA, M., and LAL, D., The record of cosmic ray intensity in the meteorites, *J. Geophys. Res.*, **66**, 3519–3531 (1961).

BAINBRIDGE, A. E., SUESS, H. E., and WANKE, H., The tritium content of three stony meteorites and one iron meteorite, *Geochim. et Cosmochim. Acta*, **26**, 471–474 (1962).

BLUE, J. W., and BLEULER, E., Radioactivity of Sc^{44}, *Phys. Rev.*, **100**, 1324–1329 (1955).

CHARABARTTY, M. M., in Progress Report, *Nuclear Chemistry Research at Carnegie Institute of Technology*, 1962–63, p. 72 (1963).

CREVECOEUR, E. H., and SCHAEFFER, O. A., Separation et mesures de Al^{26} et Be^{10} dans les meteorites, in *Radioactive Dating* I.A.E.A., pp. 335–341, Vienna, 1963.

EHMANN, W. D., and KOHMAN, T. P., Cosmic-ray-induced radioactivities in meteorites II, *Geochim. et Cosmochim. Acta*, **14**, 364–379 (1958).

FIREMAN, E. L., and DEFELICE, J., Ar^{39} and tritium in meteorites, *Geochim. et Cosmochim. Acta*, **18**, 183–192 (1960).

———, FAZIO, G., and DEFELICE, J., Argon 39, Tritium and aluminum 26 in the Farmington meteorite and their discordant exposure ages, *Trans. Am. Geophys. Union*, **44**, 89 (1963).

FISHER, D. E., and SCHAEFFER, O. A., Cosmogenic nuclear reactions in iron meteorites, *Geochim. et Cosmochim. Acta*, **20**, 5–14 (1960).

FLAGG, J. F., *Organic Reagents*, Interscience Publishers, Inc., New York (1948).

GFELLER, C., HOUTERMANS, F. G., OESCHGER, H., and SCHWARZ, U., γ–γ Koinzidenzmessung zur zerstörungsfreien Messung des Gehaltes von Meteoriten an Positronenstrahlern und γ-aktiven Isotopen, *Helv. Phys. Acta*, **34**, 466–468 (1961).

GEISS, J., HIRT, B., and OESCHGER, H., Tritium and helium concentration in meteorites, *Helv. Phys. Acta*, **33**, 590–593 (1960).

GOEL, P. S., in Progress Report, *Nuclear Chemistry Research at Carnegie Institute of Technology* 1961–1962, p. 39 (1962).

———, and KOHMAN, T. P., Cosmic-ray-exposure history of meteorites from cosmogenic Cl^{36}, in *Radioactive Dating* I.A.E.A., pp. 413–432, Vienna (1963).

HEYMANN, D., and SCHAEFFER, O. A., Personal communition (1963).

HOFFMAN, J. H., and NIER, A. O., Production of helium in iron meteorites by the action of cosmic rays, *Phys. Rev.*, **112**, 2112–2117 (1958).

———, and ———, The cosmogenic He^3 and He^4 distribution in the meteorite Carbo, *Geochim. et Cosmochim. Acta*, **17**, 32–36 (1959).

———, and ———, Cosmic-fay produced helium in the Keen Mountain and Casas Grandes meteorites, *J. Geophys. Res.*, **65**, 1063–1068 (1960).

HONDA, M., Personal communication (1962).

———, and ARNOLD, J. R., Radioactive species produced by cosmic rays in the Aroos iron meteorite, *Geochim. et Cosmochim. Acta*, **23**, 219–232 (1961).

———, and LAL, D., Spallation cross sections for long-lived radio-nuclides in iron and light nuclei, *Unpublished manuscript.*

———, SHEDLOVSKY, J. P., and ARNOLD, J. R., Radioactive species produced by cosmic rays in iron meteorites, *Geochim. et Cosmochim. Acta*, **22**, 133–154 (1961).

KAYE, J. H., Thesis, *Carnegie Institute of Technology*, NYO-8923 (1963).

MARSHALL, R. R., and HESS, D. C., Lead from troilite of the Toluca meteorite, *Geochim. et Cosmochim. Acta*, **21**, 161–164 (1961).

MAXWELL, J. A., and GRAHAM, R. P., The mercury cathode and its applications, *Chem. Revs.*, **46**, 471–498 (1950).

MAYER-BORICKE, C., BISWAS, M. M., and GENTNER, W., γ-spektroskopische Untersuchungen an Steinmeteoriten, *Z. Naturf.* **17a**, 921–924 (1962).

MURTHY, V. R., Stable isotope studies of some heavy elements in meteorites, in the volume, *Isotopic and Cosmic Chemistry* (to be published) (1963).

NININGER, H. H., The Norfork, Arkansas meteorite, an iron of witnessed fall, *Pop. Astron.*, **45**, 562–567 (1937).

PAGE, J. A., MAXWELL, J. A., and GRAHAM, R. P., Analytical applications of the mercury cathode, A review, *The Analyst*, **87**, 245–272 (1962).

PRIOR, G. T., and HEY, M. H., *Catalogue of Meteorites*, British Museum, London, 1953.

RIGHTMIRE, R. A., and KOHMAN, T. P., The half-life of long-lived Al^{26}, *Z. Naturf.*, **13a**, 847–853 (1958).

———, SIMANTON, J. R., and KOHMAN, T. P., Scheme of long-lived aluminum-26, *Phys. Rev.*, **113**, 1069–1077 (1959).

SAMMET, F., and HERR, W., Studies on the cosmic-ray produced nuclides Be^{10}, Al^{26} and Cl^{36} in iron meteorites, in *Radioactive Dating* I.A.E.A., pp. 343–354, Vienna (1963).

SANDELL, E. B., *Colorimetric Methods for the Determination of Traces of Metals*, Interscience Publishers, New York, 1944.

SIGNER, P., Unpublished data (1962).

———, and NIER, A. O., The distribution of cosmic-ray produced rare gases in meteorites, *J. Geophys. Res.*, **65**, 2947–2964 (1960).

———, and ———, The measurement and interpretation of rare gas concentrations in iron meteorites, in *Researches on Meteorites*, edited by C. B. Moore, pp. 7–35, John Wiley & Sons, Inc., New York, 1962.

SIMPSON, C. T., and CHANDLEE, G. C., *p*-Hydroxyphenylarsonic acid as a reagent for titanium and zirconium, *Ind. Eng. Chem. Anal. Ed.*, **10**, 642–642 (1938).

UREY, H. C., and MURTHY, V. R., Isotopic abundance variations in meteorites (abstract), *Science*, **140**, 385–386 (1963).

VILCSEK, E., and WANKE, H., Das Strahlungsalter der Eisenmeteorite aus Chlor-36-Messungen, *Z. Naturf.*, **16a**, 379–384 (1961).

———, and ———, Cosmic-ray-exposure ages and terrestrial ages of stone and iron meteorites derived from Cl^{36} and Ar^{30} measurements, in *Radioactive Dating* I.A.E.A., pp. 381–393, Vienna (1963).

VISTE, E., and ANDERS, E., Cosmic-ray-exposure history of tektites, *J. Geophys. Res.*, **67**, 2913–2929 (1962).

VOSHAGE, H., and HINTENBERGER, H., Cosmic-ray-exposure ages of iron meteorites, in *Radioactive Dating* I.A.E.A., pp. 367–379, Vienna (1963).

WANKE, H., Scandium 45 als Reaktionsprodukt der Höhenstrahlung in Eisenmeteoriten II, *Z. Naturf.*, **15a**, 953–964 (1960).

———, and HINTENBERGER, H., Helium und Neon als Reaktionsprodukt der Höhenstrahlung in Eisenmeteoriten, *Z. Naturf.*, **13a**, 895–897 (1958).

COHENITE AS A PRESSURE INDICATOR IN IRON METEORITES?

MICHAEL E. LIPSCHUTZ*

Goddard Space Flight Center

AND

EDWARD ANDERS

Enrico Fermi Institute for Nuclear Studies, University of Chicago

Cohenite in iron meteorites must have formed below the α–γ transformation temperature, probably at 680°±20°C. Laboratory experiments on the decomposition of Fe_3C at high pressures fail to show the stabilization predicted by Ringwood, and it has therefore been necessary to revise the high-pressure portion of the Fe–C diagram. If iron meteorites had cooled slowly at pressures of 30–60 kb (Ringwood and Kaufman, 1962), cohenite should have decomposed in a matter of hours to graphite or diamond, at temperatures below 640 to 565°C. The times actually available for decomposition were longer by factors of 10^{10} to 10^{11}, and one must therefore assume that nucleation was inhibited by a factor of this order.

If the iron meteorites had originated at low pressures, precisely the same assumption (nucleation inhibited by a factor of 10^{10}–10^{11}) must be made. Hence there is no longer any reason to attribute the survival of cohenite to high pressures. Laboratory experiments show that the decomposition of cementite at 680°C and 1 atm can indeed be inhibited by factors $>10^5$. It does not seem unreasonable that meteoritic cohenite, having a much higher degree of crystal perfection, should be inhibited in its decomposition to an even greater extent.

Ringwood (1960) has argued that the presence of cohenite [$(Fe, Ni)_3C$] in iron meteorites implies high pressures in the meteorite parent bodies, since the pure compound Fe_3C is thermodynamically unstable at low temperatures and pressures. In a later paper (Ringwood and Seabrook, 1962), he showed that cohenite had a very low phosphorus content (0.10–0.49 wt pct), and was therefore not stabilized by phosphorus, as we had suggested (Lipschutz and Anders, 1961a). He also noted that individual cohenite crystals heated to 750°–950°C decompose in a matter of hours, and argued that this evidence supported his earlier conclusion.

Although we are impressed by Ringwood and Seabrook's evidence that the phosphorus content of cohenite is too low to stabilize it to a significant extent, we find ourselves unable to accept their conclusion: that it was stabilized by high pressure. In this paper, we want to direct attention to certain factors that have not received sufficient attention in previous papers on the subject.

1. FORMATION TEMPERATURE OF COHENITE

Let us first discuss the formation of cohenite in terms of the binary Fe–C phase diagram (fig. 1), and then make the necessary allowances for the ternary Fe–Ni–C system. Even cohenite-rich iron meteorites generally have carbon contents no higher than 0.4–0.6 weight percent, that is, less than the α-γ eutectoid composition (0.80%C). For structural reasons, carbon is much more readily soluble in face-centered cubic γ-phase than in body-centered α-phase. Hence a hypo-eutectoid iron-carbon alloy of the carbon content of Canyon Diablo (0.6 wt pct; Brentnall and Axon, 1962) cooling from high temperatures

*On active duty with the U.S. Army.

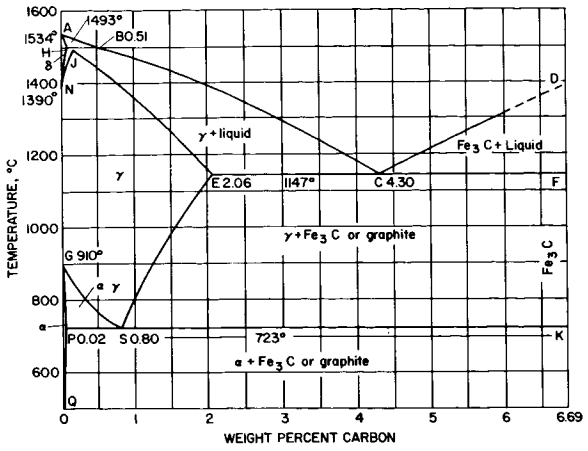


FIGURE 1.—Iron-carbon phase diagram (Hansen, 1958). When a hypoeutectoid alloy (<0.80 weight percent C) crosses the line GS on cooling, α-iron precipitates until the residual γ-phase has reached eutectoid composition (point S). Only then, at 723°C, does Fe_3C appear.

remains homogeneous γ-phase down to about 750°C; then, carbon-poor α-phase begins to precipitate, and the residual γ-phase becomes progressively richer in carbon until the eutectoid temperature (723°C) has been reached. At this point, the alloy breaks down into α-phase and metastable cementite.* Although graphite is the stable reaction product, metastable cementite is favored on kinetic grounds. Thus the steel industry is able to produce cementite in multi-megaton quantities annually.

In the presence of nickel, the situation changes slightly (fig. 2). Nickel stabilizes the γ-phase, causing the γ-field in fig. 1 to extend to lower temperatures and perhaps to higher carbon contents. On cooling, the carbon again remains in the γ-phase (taenite), while carbon- and nickel-poor α-phase (kamacite) precipitates. But there is no longer a eutectoid temperature at which taenite breaks down into kamacite and cohenite. Even below 723°C, carbon-rich taenite persists, stabilized by its nickel content. As the kamacite grows at the expense of taenite, nickel and carbon concentrate in the taenite, until the alloy becomes saturated with carbon. At this temperature, well below 723°C, cohenite makes its first appearance: cohenite crystals nucleate at the kamacite-

*We shall use the terms *cementite* for the pure compound Fe_3C, and *cohenite* for the meteoritic mineral $(Fe, Ni)_3C$, containing about 2% Ni.

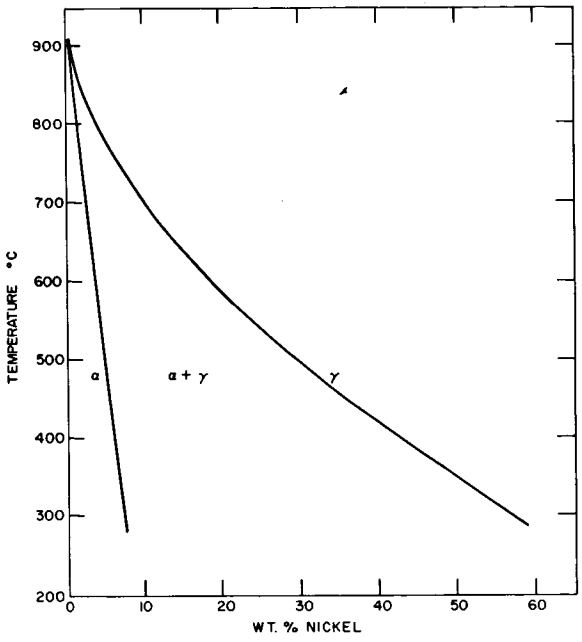


FIGURE 2.—Iron-nickel phase diagram (Owen and Liu, 1949). An alloy of 7.5% Ni remains homogeneous taenite down to 740°C, where kamacite first appears. On further cooling, carbon remains in the taenite until its solubility limit is reached. At this point, Fe_3C precipitates.

taenite interface, and continue to grow while more and more carbon is being rejected by the shrinking taenite lamellae. This continued slow growth over a substantial temperature range permits cohenite crystals in Fe–Ni–C alloys to grow much larger than the microscopic carbide particles in pearlite, the Fe_3C/α–Fe eutectoid, which precipitates in Fe-C alloys on a very much shorter time scale and at a far higher degree of supersaturation. It does not seem necessary to invoke metastability (Ringwood, 1960) to account for this size difference.

The actual temperature at which cohenite first appeared can be estimated from Brentnall and Axon's (1962) data. They reheated cohenite-bearing samples of Canyon Diablo at atmospheric pressure, and found that cohenite crystals redissolved completely in their kamacite matrix at 700°C in 7 days, but remained unchanged, except for a thin ferrite halo, at 650°C. These experiments do not quite duplicate the formation conditions of cohenite, since the cohenite is now in contact with kamacite of 7% Ni, rather than the

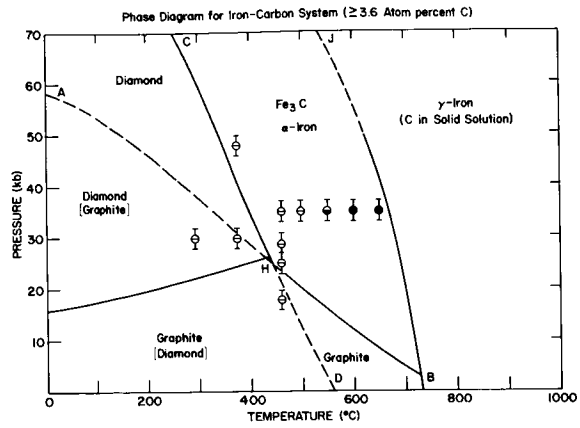


FIGURE 3.—Phase diagram for iron-carbon system ($\leq$0.8 weight percent C). Adapted from Lipschutz and Anders (1961a), but with Darken and Gurry's (1951) thermodynamic data for Fe_3C substituted for the original values. Metastable phases resulting from the decomposition of Fe_3C are enclosed in brackets. Cementite is stable only in the field CHBJ, but it can form as a metastable phase everywhere to the left of the γ-iron field.

The points represent experimental measurements on the decomposition of Fe_3C (Lipschutz and Anders, 1961b). Filled circles: extensive graphitization; half-filled circles: partial graphitization; circles with horizontal bars: amorphous carbon recovered after dissolving sample in acid, but this carbon may have formed during dissolution rather than during heat treatment (see text). Decomposition of cementite in its apparent stability field shows that the true stability field of cementite must be smaller than that indicated in this diagram.

original taenite of perhaps 8–11% Ni. Nevertheless, the error thus introduced is negligible for our purposes, and we can conclude, for Canyon Diablo at least, that the formation temperature of cohenite (at 1 atm) lies between 650° and 700°C, most probably near 680°C. Although this figure is uncertain by perhaps ±20°C, the conclusions reached in this paper are quite insensitive to an error of this magnitude.

2. DECOMPOSITION OF COHENITE TO GRAPHITE OR DIAMOND

Cementite is thermodynamically unstable at low temperatures and pressures:

$Fe_3C \rightarrow 3Fe + C$ (graphite) $\Delta F^\circ_{298} = -4.8$ kcal/mol (Darken and Gurry 1951).

However, this reaction proceeds with a large, positive volume change ($\Delta V_{298} = +3.419$ cc/mol). Ringwood (1960) therefore pointed out that the negative free energy would be canceled by the $P\Delta V$ term at sufficiently high pressures.

However, we have emphasized in several of our papers (Lipschutz and Anders, 1961 a, b) that diamond, rather than graphite will be the stable reaction product at pressures above the diamond-graphite transformation line. Consequently, the following reaction must also be considered:

$Fe_3C \rightarrow 3Fe + C$ (diamond) $\Delta F^\circ_{298} = -4.1$ kcal/mol; $\Delta V_{298} = +1.500$ cc/mol

If the pressures needed to stabilize cementite are calculated for both reactions, the following phase diagram results (fig. 3).* From this diagram it is seen that if iron meteorites had cooled isobarically at 30–60 kb (Ringwood, 1960, 1961; Ringwood and Kaufman, 1962), *cohenite should have decomposed to diamond* and kamacite at temperatures of 420–290°C. To explain the survival of cohenite, it is necessary to assume that the reaction rate was very slow at these temperatures. We shall return to this point later on.

However, the decomposition of cementite at high pressures has been studied experimentally (Lipschutz and Anders, 1961b), and the results do not show the stabilization predicted by Ringwood on the basis of Darken and Gurry's data. At temperatures of 550° to 650°C and pressures of 35 kb (fig. 3) cementite decomposed, in a matter of days, to crystalline graphite.†

Clearly, the stability field of Fe_3C is much smaller than indicated by Darken and Gurry's data. To bring this dicegram into accord with the

*In this diagram we have used Ringwood's choice of thermodynamic data ($\Delta F^\circ_{298} = -4.76$ kcal/mol, Darken and Gurry, 1951), instead of the value from NBS Circular 500 ($\Delta F^\circ_{298} = -3.5$ kcal/mole) used in our earlier paper. This change shifts the intercept of lines AB and CD with the abscissa to higher temperatures, but leaves their slopes virtually unchanged. The γ–α eutectoid temperature at high pressures was taken from Claussen's (1963) paper.

†At temperatures *below* 550°C, only amorphous carbon rather than crystalline graphite was found after dissolving the sample in HCl. Although control samples of cementite not subjected to heat treatment left only a minute trace of residue after HCl treatment, we are not certain that the amorphous carbon actually formed during heat treatment rather than during dissolution in acid. Perhaps the heat treatment induced some subtle changes in the cementite that favored the formation of carbon over hydrocarbons in its reaction with acid.

experimental data, it is necessary to shift lines AB and CD to the right, by an amount corresponding to an error of about 2.3 kcal/mole in the standard enthalpy (fig. 4). Such an error is entirely consistent with the uncertainty in Darken and Gurry's extrapolation, since the data from which they derived $\Delta H_{298}°$ show several conflicting trends (fig. 5 in their paper).

This change in the diagram may seem arbitrary, but it must be remembered that the point under discussion is the stability of Fe_3C at high pressures. Experimental data on the behavior of Fe_3C at high pressures must therefore be given greater weight than stability relations derived by indirect methods.

Now we are in a position to estimate the formation and decomposition temperature of cohenite, at any pressure. Three cases will be considered.

A. Cooling of iron meteorites at pressures no greater than 2.7 kb (the central pressure in Ceres, the largest asteroid). As pointed out above, the formation temperature of cohenite will be ~680°C, about 40° below the α–Fe/γ–Fe eutectoid temperature (fig. 4). Since the cohenite is formed as a metastable phase, its decomposition temperature will also be 680°C, and the expected decomposition product, graphite.

B. Cooling of iron meteorites at 30 kb, the minimum pressure which Ringwood and Kaufman (1962) consider suitable for the formation of the Widmanstätten pattern. At this pressure, the α–Fe/γ–Fe eutectoid temperature is depressed to ~680°C (fig. 4) and if we once again assume that cohenite formed 40° below the eutectoid, its formation temperature will be 640°C. Even at this high pressure it cannot form as a stable phase. Just as in case A, it will form as a metastable phase, and its decomposition temperature will hence equal its formation temperature, 640°C.

C. Cooling of iron meteorites at 60 kb, the upper limit of Ringwood and Kaufman's pressure range. Proceeding as in case B, we can estimate the formation temperature of cohenite as 580°. It

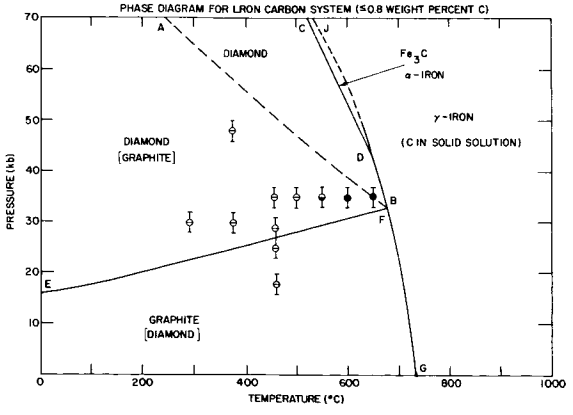


FIGURE 4.—Revised phase diagram for iron-carbon system (≤0.8 weight percent C). As in figure 3, with lines AB and CD shifted to the right in order to account for the observed instability of Fe_3C at high pressures. The stability field of cementite has now contracted to the region CJD.

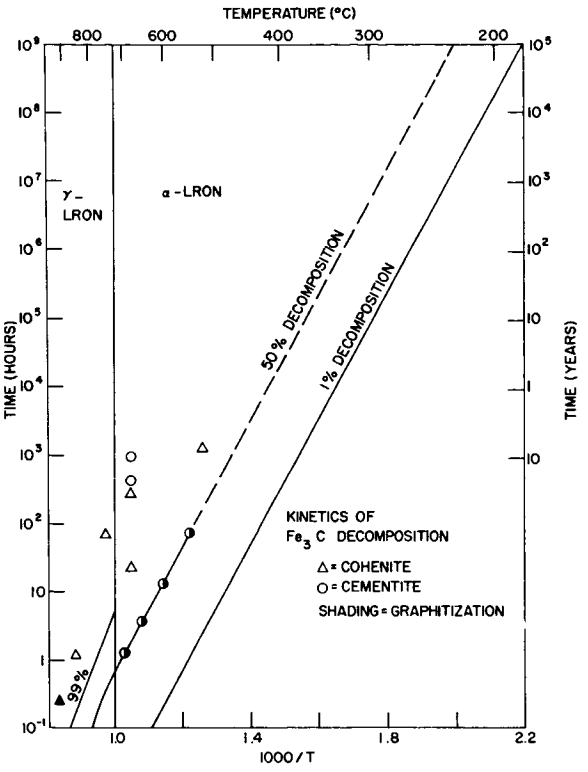


FIGURE 5.—Rate of decomposition of cohenite and cementite. The points through which the 50% graphitization line has been drawn are from Klein; the three points in the γ-field, from Ringwood and Seabrook; and the remaining points, from this study.

Cementite samples heated at 680°C for 970 hrs. failed to show any graphitization whatsoever, although similar samples with pre-formed nucleation centers show 1% graphitization in 0.03 hrs. Cohenite samples decompose somewhat more readily, but the graphitization is localized at cracks.

is now formed as a stable phase, however, and does not become unstable with respect to diamond until the temperature has fallen to 565°C.

It is now clear that even at the highest pressures considered by Ringwood, cohenite is not "stabilized" permanently. On further cooling, cohenite becomes unstable again, and should decompose to graphite or diamond, as shown below:

Pressure (kb)	0	30	60
Decomposition temperature (°C)	680	640	565
Decomposition product	graphite	graphite	diamond

At these temperatures the Widmanstätten pattern had not yet developed. Since its formation requires very slow cooling, on a time scale of millions of years, one must ask whether the survival of cohenite at these times and temperatures is consistent with laboratory data on the kinetics of cementite decomposition.

3. RATE OF DECOMPOSITION OF COHENITE

Ringwood and Seabrook attempted to determine this rate by heating individual cohenite crystals to temperatures of 750°–950°C. Unfortunately, these results are not germane to the problem. A single cohenite crystal heated by itself represents a hypereutectoid alloy with 6.7% C; when heated above the α–Fe/γ–Fe eutectoid temperature, it must decompose to graphite and γ-iron saturated with carbon. Meteorites, on the other hand, are hypoeutectoid alloys ($\geq$0.8% C). As noted above, cohenite will not appear in these alloys until ~680°C, when the taenite has become saturated with carbon. At and *below* its formation temperature, cohenite can only decompose to α-iron* and carbon. *Above* its formation temperature, cohenite will redissolve in the surrounding kamacite, as soon as the α–γ eutectoid temperature for an Fe–C alloy with 7% Ni has been reached. But this situation is artificial, and has no bearing on the survival of cohenite in the meteorite parent bodies, under conditions of monotonic cooling.

Hence it is essential to determine the decomposition rate of Fe_3C *below* the eutectoid temperature. A thorough study of this subject was made by Klein (1934). He found that the decomposition of Fe_3C in the temperature range 550°C to 700°C obeys first-order kinetics (with a short induction period attributed to diffusion of carbon from Fe_3C to graphite nuclei). The activation energy is 42 kcal/mole.

Figure 5 summarizes the available data on the decomposition rate of Fe_3C. Two straight lines with slopes of 42 kcal/mole have been drawn to represent Klein's data for 50% and 1% graphitization. The 50% line has been extended into the γ-region, with an activation energy of 58 kcal/mole (Kunyavskii and Khudokormov, 1960). Ringwood and Seabrook's points, corresponding to "complete" graphitization, lie somewhat above the 99% line, but the disagreement is not serious in view of the uncertainty in defining "complete" graphitization.

Of course, only points in the α-region are pertinent to the problem. In order to provide data on the behavior of cohenite in this region, we heated cohenite-bearing Odessa samples and a cementite (white cast iron) sample (3%C) to 680°C for varying lengths of time. Nickel diffusion is negligible at the times and temperatures involved, and the results on both cohenite and cementite can therefore by discussed in terms of the binary Fe–C diagram. According to Klein's data, the samples should have been 50% graphitized after 1.8 hours. Actually, the cementite sample showed no decomposition whatsoever ($\ll$1%) even after 970 hours, while the cohenite, after heating for 355 hours, showed partial (~10%) graphitization, localized at cracks.

That cracks, by providing nucleation sites, accelerate graphitization, had already been noted by Brentnall and Axon (1962). They found that cracked cohenite lamellae in Canyon Diablo graphitized readily at 708°C in 24 hours, but that a single uncracked cohenite crystal merely dissolved in the surrounding metal, without showing any trace of graphitization. We were able to produce the same contrasting behavior in the cementite sample, by crushing it at room temperature with a static force of 10,000 kg/cm².* During heat treat-

*This is a consequence of the low Ni-content (~2%) of cohenite. The Ni–Fe alloy formed in its decomposition will be low in Ni, and according to either figure 1 or 2, this alloy can exist only in the α-form at or below the cohenite formation temperature.

*This experiment was suggested by Prof. O. J. Kleppa of the University of Chicago.

ment, the crushed sample graphitized at cracks (fig. 6), in contrast to the uncrushed sample. Cohenite, on the other hand, decomposed at cracks to about the same degree, regardless of whether the sample had been crushed or not. Presumably sufficient cracks had been produced in the cohenite during breakup of the meteorite parent body, or during impact with the earth, so that further crushing did not cause a significant increase in nucleation centers. However, there is every reason to believe that the cohenite crystals were free of cracks until the breakup of the meteorite parent body. Estimates of their decomposition rate should hence be based on uncracked material, free from shock effects.

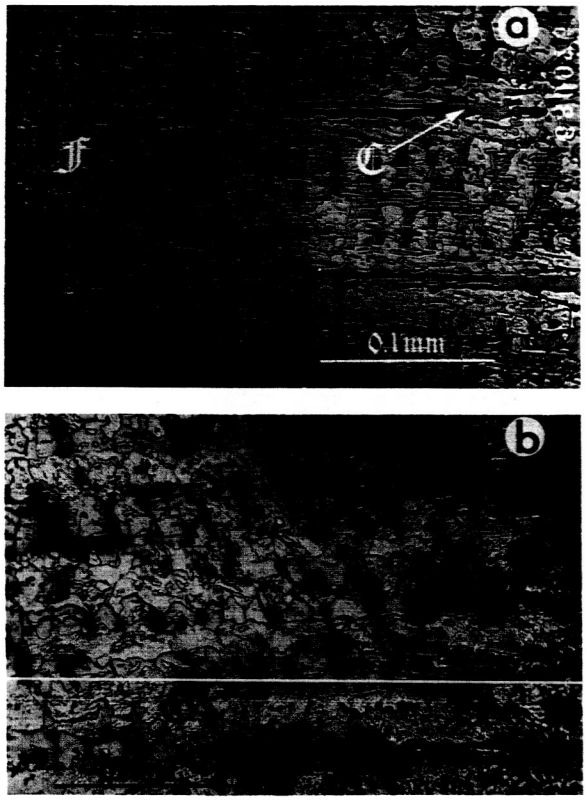


FIGURE 6.—Effect of crystal perfection on graphitization rate. An untreated cementite sample (a) showed no graphitization after 355 hr at 680°C, while an identical sample (b) crushed with a static force of 10,000 kg/cm² before heat treatment graphitized readily at cracks. Although the area photographed was rather more extensively cracked and graphitized than the sample as a whole, a measurable degree of graphitization was visible throughout the specimen. C = cementite; F = ferrite (α-iron); G = graphite.

As little as 1% graphitization would have been easily noticeable in the cementite sample heated at 680°C for 970 hrs. Actually, not a trace of graphite was found, although 1% graphitization should have taken only 3×10^{-2} hours, according to Klein. Hence the observed rate is a factor of $\gg3\times10^{4}$ slower than that expected from Klein's kinetic study!

But this disagreement is not entirely unexpected. As pointed out by Ringwood (1960), Klein (1934), and many others, the graphitization of cementite requires nucleation centers: graphite particles, other impurities, or crystal imperfections. Normally such centers are produced by heating the alloy in the γ-region above the eutectoid temperature. If this step is omitted, graphitization in the α-region is inhibited. Klein remarks that white cast iron of the same composition as his samples, but without nucleation centers . . . "did not decompose at temperatures up to 700°C, even after heating for many days".

Now we can return to the central question of this paper: the survival of cohenite in meteorites. We saw above that the stability field of cohenite is quite narrow even at high pressures, and that it should decompose to diamond or graphite on cooling. We must now determine whether the rate of graphitization is slow enough at the decomposition temperature to permit its survival.

Let us consider the three cases for which we had previously determined formation and decomposition temperatures of cohenite. The time during which cohenite finds itself at elevated temperatures and hence is in danger of decomposition, is determined by the time needed to form the Widmanstätten pattern. At low pressures, a cooling rate of 160° per million years seems to be required (Wood, 1963). At higher pressures, the γ–α transformation temperature is depressed; to allow the necessary diffusion of nickel to take place, a slower cooling rate and longer time must be assumed. Ringwood and Kaufman (1962) estimate a formation time of 10^{8} years. This is equivalent to an average cooling rate of 11°C per million years, if the entire temperature drop from 1500°C to 400°C is assumed to have taken place in this interval.

Once the temperature has fallen below the decomposition temperature of cohenite, $\bar{T}_d$, the

TABLE I.—*Comparison of Cohenite Decomposition Rate with Times Available for Decomposition*

Pressure (kb)	Cohenite		Predicted time for 1% decomposition (hrs.)	Effective time at decomposition temperature (yrs.)	Inhibition factor
	Formation temperature (°C)	Decomposition temperature (°C)			
<2.7	680°	680°	0.026	2.4×10^{5}	8.1×10^{10}
30	640°	640°	0.068	2.9×10^{6}	3.7×10^{11}
60	580°	565°	0.56	1.5×10^{6}	2.4×10^{10}

reaction will proceed continuously. Owing to the exponential decrease of the rate constant with temperature, most of the decomposition will take place near T_d. In fact, one-half the total decomposition will take place in the first half-life, defined as the time during which the temperature decrease causes the rate to drop to one-half its initial value. It follows that the total decomposition, during cooling from T_d to some final temperature where the rate becomes negligibly slow, will be exactly the same as if the sample had been held at T_d for an "effective time" equal to one mean life ($t_{1/2}/0.693$). These effective times, or times available for decomposition, may be compared directly with the decomposition times predicted from laboratory experiments (table I). The last column gives the inhibition factor: The ratio of the available decomposition time to the time required for 1% decomposition in the laboratory.

Clearly, cohenite has failed to make use of the opportunity to decompose, although the available times were 10^{10}–10^{11} times longer than those required for 1% decomposition. This is equally true at high and at low pressures, and it is no longer possible to attribute the survival of cohenite to high pressures.

In making this comparison, we have tacitly assumed that the decomposition of cohenite to diamond will proceed at about the same rate as its decomposition to graphite. This assumption is probably justified. The rate-determining step in the decomposition reaction is probably not the diffusion of carbon atoms through α-iron (for which the activation energy is only 20.1 kcal/mole, Wert 1950), but the rearrangement of iron atoms in the cementite lattice to the body-centered cubic structure of α-iron. The activation energy for the latter process should be the same reagrdless of whether carbon is eliminated as graphite or diamond.

One might be tempted to rescue the high-pressure hypothesis by some special assumption. Could it be that the decomposition rate of Fe_3C at high pressures is much lower than the rate at 1 atm upon which these comparisons are based? Fortunately, there exist laboratory data on this question. Cementite heated under 35 kb pressure graphitized extensively after 48 hrs at 600°C and 69 hrs at 650°C (Lipschutz and Anders, 1961b). The half-times for graphitization at these temperatures, as given by Klein's data, are 13 hrs and 3.3 hrs. Hence there is no significant difference between the high-pressure and low-pressure decomposition rates in the laboratory.

It does not seem possible to attribute the survival of cohenite to stabilization by trace elements. The present phosphorus content is evidently too low, and nickel, although present in greater amounts (Brown and Lipschutz, 1963), can at best stabilize it by 0.20 cal $mole^{-1}$ deg^{-1} due to the entropy of mixing.

Only one possibility seems to remain: slow nucleation. Our own experiments show that absence of nucleation centers inhibits graphitization of cementite at 680°C by a factor of $>3\times10^{4}$. Other authors (Hickley and Quarrell, 1954) have observed factors of $>10^{5}$. In meteorites, factors of 10^{10}–10^{11} are needed, and while we do not expect to be able to prove experimentally within our lifetimes that such factors are realistic, we want to point out a few reasons why nucleation

might be slower in meteorite parent bodies than in the laboratory.

Nucleation in commercial alloys is usually enhanced by Si or Al; apparently, because the $(1\bar{1}00)$ planes of graphite nucleate readily on the (110) planes of cristobalite and γ-alumina (Hickley and Quarrell, 1954). In support of this mechanism, these authors showed that the rate of graphitization was very slow *in vacuo*, but increased with increasing partial pressure of oxygen. Cohenite-bearing iron meteorites have undoubtedly formed under an exceedingly low partial pressure of oxygen, and they do not contain cristobalite or γ-alumina. Any other oxides present are usually localized in troilite nodules where they are out of contact with the cohenite. Other types of nucleation centers, such as dislocations, cracks, and impurities, are likely to be much rarer in cohenite crystals grown by slow cooling during 10^5–10^0 years than in commercial alloys cooled within minutes or hours. The activation energy for graphite nucleation on cristobalite is 79 kcal/mole between 800 and 1000°C (Burke and Owen, 1954), but for less favorable nucleation centers the activation energy may well be considerably higher. Finally, the free-energy difference between cementite and graphite is so small near its formation temperature, that it may be more than offset by the surface energy, which will favor millimeter-sized cohenite crystals over Angström-sized graphite nuclei. To sum up the matter: If relatively imperfect cementite crystals heat-treated in the laboratory show inhibition by factors of $>10^5$, is it entirely unreasonably to suppose that the larger and much more perfect meteoritic cohenite crystals were inhibited by factors of 10^{10}–10^{11}?

Ringwood (1960) has argued that iron meteorites, having cooled slowly through the temperature range 1000°–700°C where "graphite nuclei are readily and quickly formed", should provide a favorable medium for the decomposition of cohenite. Actually, the opposite is more likely to be true. Hypoeutectoid alloys will not form graphite nuclei, since they can retain all their carbon in solution down to 680°C, where cohenite first forms. Hypereutectoid alloys, on the other hand, will precipitate graphite at temperatures above the eutectoid, but since carbon can diffuse freely at the enormously long cooling times available, this graphite will segregate at the periphery of troilite nodules, rather than remaining dispersed in the metal, where it might later provide nucleation centers for the decomposition of cohenite. This is in accord with observation: graphite is never found in the metal but only in and around nodules of other minor constituents.

We admit that the notion of reaction inhibition by a factor of 10^{11} grossly violates one of the most basic premises in geology and meteoritics: That times were long enough for equilibrium to be reached. Every effort should therefore be made to find alternative explanations of the survival of cohenite. However, one point may already be stated with emphasis: *cohenite initially stabilized by high pressures becomes unstable again on cooling, and cannot survive, unless its decomposition is inhibited by a factor of* 10^{11}. Since inhibition by the same factor is required at low pressures, there is no longer any reason to regard cohenite as a "pressure indicator" in iron meteorites.

ACKNOWLEDGMENTS

We thank Mr. Pedro D. Sarmiento of the Materials Research and Development Section, Goddard Space Flight Center, for the preparation of the photomicrographs. This work was supported in part by the National Science Foundation (Grant Number NSF–G14298) and by the U.S. Atomic Energy Commission (Contract AT(11–1) 382).

REFERENCES

BRENTNALL, W. D., and AXON, H. J., The response of Canyon Diablo meteorite to heat treatment, *J. Iron Steel Inst.* (London), **200,** 947–955 (1962).

BROWN, J., and LIPSCHUTZ, M. E., Unpublished data (1963).

BURKE, J., and OWEN, W. S., Kinetics of First-Stage Graphitization in Iron–Carbon–Silicon Alloys, *J. Iron Steel Inst.* (London), **176,** 147–155 (1954).

CLAUSSEN, W. F., Alpha–Gamma Transformation in Iron Alloys—Calibration of Pressure by Duplex Differential Thermal Conductivity Analysis, Report No. 62–RL–3088M (1962).

DARKEN, L. S., and GURRY, R. W., Free energy of formation of cementite and the solubility of cementite in austenite, *Trans. AIME,* **191,** 1015–1018 (1951).

HANSEN, M., *Constitution of Binary Alloys*, McGraw-Hill, New York (1958).

HICKLEY, R. H., and QUARRELL, A. G., The graphitization of steel at subcritical temperatures, *J. Iron Steel Inst.* (London), **178,** 337–346 (1954).

KLEIN, E. H., Der zeitliche Verlauf des Zementitzerfalls im Gusseisen, *Stahl und Eisen,* **54,** 827–830 (1934).

KUNYAVSKII, M. N., and KHUDOKORMOV, V. N., Decomposition of cementite in white cast iron, *Izvestiya Vysshikh Ucheb. Zavedenii Mashinostroeniye,* **1960,** 17–23 (1960).

LIPSCHUTZ, M. E., and ANDERS, E., The Record in the Meteorites–IV. Origin of diamonds in iron meteorites, *Geochim. et Cosmochim. Acta,* **24,** 83–105 (1961a).

———, and ———, On the mechanism of diamond formation, *Science,* **134,** 2095–2099 (1961b).

OWEN, E. A., and LIU, Y. H., Further X-ray study of the equilibrium diagram of the iron–nickel system, *J. Iron Steel Inst.* (London), **163,** 132–137 (1949).

RINGWOOD, A. E., Cohenite as a pressure indicator in iron meteorites, *Geochim. et Cosmochim. Acta,* **20,** 155–158 (1960).

———, Chemical and genetic relationships among meteorites, *Geochim. et Cosmochim. Acta,* **24,** 159–197 (1961).

———, and KAUFMAN, L., The influence of high pressure on transformation equilibria in iron meteorites, *Geochim. et Cosmochim. Acta,* **26,** 999–1010 (1962).

———, and SEABROOK, M., Cohenite as a pressure indicator in iron meteorites II, *Geochim. et Cosmochim. Acta,* **26,** 507–510 (1962).

WERT, C. A., Diffusion coefficient of C in α-Iron, *Phys. Rev.,* **79,** 601–605 (1950).

WOOD, J. A., Unpublished data (1963).

EVOLUTION OF O STARS. I. HYDROGEN-BURNING

RICHARD STOTHERS

Harvard College Observatory and Goddard Space Flight Center

The evolution of an O-type star of 30 $M\odot$ is considered from its initial main-sequence state to the point where the hydrogen content of the core has decreased from 70% to 7% by weight. The O stars are characterized by convective instability in the envelope and the full CNO-cycle of energy generation. Radiation pressure is important throughout the star, and electron scattering dominates the opacity. The inhomogeneous models consist of four zones: (I) a radiative envelope retaining the initial composition, (II) an outer semiconvective zone with a continuously varying composition that has a discontinuity at the interface with (III) an inner radiative zone also with a continuously varying composition, and (IV) a homogeneous convective core. Due to increasing radiation pressure, the semiconvective zone grows more rapidly outward than inward, but produces little effect on the important stellar quantities. The mass fraction of the core decreases from 0.60 to 0.34 and the mean hydrogen content of the star gets 40 per cent depleted. The age, 4.7 million years, is shorter than that for other sequences computed for stars of mass $\approx 30\ M\odot$ because of a higher luminosity. This is mainly the result of choosing a higher mean molecular weight and CNO content.

INTRODUCTION

It is believed that the early-type spectra of the H–R diagrams of very luminous OB clusters and associations can be explained by a main sequence of rapidly evolving massive stars. Schwarzschild and Härm (1959) showed that the upper end of the main sequence of stable stars may lie at about 60 $M\odot$; this should correspond to the hottest of normal O5 spectra.

Sakashita, Ôno, and Hayashi (1959) have studied the evolution of a star with mass 15.6 $M\odot$ and initial spectrum B0, and found a growing convective instability in the envelope. This instability does not occur in stars with mass $\lesssim 10\ M\odot$ (Schwarzschild, and Härm, 1958) and remains negligible up to about the mass they chose. For more massive stars, however, it must be taken into account. Increasingly important is the role of O^{16} in the hydrogen-burning reactions. At temperatures above 3×10^7 °K (the central temperature of the main-sequence model for a star of 15.6 $M\odot$) the CNO-cycle replaces the CN-cycle. Thus we find characteristic of the O stars: (1) convective instability in the envelope and (2) the complete CNO-cycle.

The massive stars with $M \lesssim 10\ M\odot$ have been discussed by several workers (Taylor, 1954; Kushwaha, 1957; Blacker, 1958; Henyey, LeLevier, and Levée, 1959; Hoyle, 1960). These stars are characterized by a *shrinking* of their convective regions. Their evolution is perhaps best described by Kushwaha (1957).

In this paper we propose to study in detail the early evolution of a star with mass lying in the region of the O stars, heretofore studied only approximately (Blackler, 1958; Schwarzschild, and Härm, 1958; Henyey et al., 1959; Hoyle, 1960). We choose a mass of 30 $M\odot$ and consider the evolution to the point where the hydrogen content of the core drops to 10% of its initial value.

ASSUMPTIONS

The usual general assumptions are made.

1. Mass loss is neglected for the early evolutionary phases, following, for example, the summary by Deutsch (1960) and the note by Huang, Milligan, and Stecher (1962). Were mass loss significant, it should be easily observable (Mustel, 1959), since even among the late-type giants where the surface gravity is much lower, mass loss is not always observed.

2. Rotation is neglected. Although more massive stars on the main sequence are observed to rotate faster, supergiants have little or no rotation. Theoretical justification exists for the necessary braking during evolution (see, e.g., Huang and Struve, 1960). Even rotation during the main-sequence phase probably produces little mixing, since the observed Trumpler effect in galactic clusters suggests evolution without mixing for many of the early-type stars (Deutsch, 1960). In fact, theory predicts longtime scales of mixing, and Mestel (1959) has calculated that there will be mixing between the core and envelope only if the star is near disruption. By neglecting rotation, we also neglect departures from spherical symmetry.

3. No convective overshooting and consequent mixing occurs between any of the convective-radiative interfaces in the star. Sakashita and Hayashi (1961) have discussed this assumption.

4. The turbulent mixing in the core is complete and maintains a steady homogeneity of composition.

Calculation of the structure of the star is considerably simplified by the following additional conditions.

1. The surface temperature of the O and early B stars is so high that the outer convection zone is negligible. This means that we may consider the envelope as in radiative equilibrium out to the boundary.

2. The convective core contains such a large fraction of the stellar mass that effectively all of the nuclear energy generation occurs there.

3. Degeneracy does not set in even at the center. If we may extrapolate the results of Hayashi and Cameron (1962) for a star of 15.6 $M\odot$, the gas will remain nondegenerate at least up to the carbon-burning phase.

4. The opacity may be assumed with good accuracy to depend only on electron scattering (cf., Hayashi and Cameron, 1962).

Appreciable structural complications occur in very massive stars, however. Radiation pressure must be fully taken into account throughout, and the inner part of the envelope becomes convectively unstable. Thus, simplification of the physics is offset by increasing structural difficulties.

GENERAL STRUCTURE

The internal structure of very massive stars is now fairly well understood, as the result of the accumulated efforts of several workers over the past few years. Following the work of Taylor (1954) and Kushwaha (1957) on stars of mass less than 10 $M\odot$, in which a shrinking convective core leaves behind a radiative intermediate zone with continuous gradient of mean molecular weight, Schwarzschild and Härm (1958) showed that in stars of higher mass, radiation pressure makes the region just outside the core convectively unstable. Convection mixes the composition until the unstable zone attains convective neutrality. The resulting distribution of chemical composition was determined by equating the radiative and adiabatic temperature gradients.

Sakashita et al. (1959) pointed out that the existence of a gradient of mean molecular weight will in turn act to suppress the convective motion in the semiconvective zone. Therefore, a term including this gradient must be added to the adiabatic temperature gradient. They also showed that the predominance of electron scattering in the opacity requires the region at the boundary of the core to be in radiative equilibrium, in order to preserve continuity of the radiative energy flux. Assuming, however, that the convective motion in the intermediate zone is sufficiently effective, Sakashita and Hayashi (1959) discussed the case of a star of 46.8 $M\odot$, in which the purely radiative zone does not appear because of overshooting mass motion. This may be true in the case where hydrogen has not been extensively depleted in the core, but eventually continuity

of the radiative flux at the boundary of the core will require a radiative zone to form inside the semiconvective zone.

Sakashita and Hayashi (1961) improved their model by dividing it into four zones: (I) a homogeneous radiative envelope retaining the initial composition, (II) a semi-convective outer intermediate zone, (III) a radiative inner intermediate zone, and (IV) a homogeneous convective core. In this paper their notation will be followed. We introduce the following subscripts: 0 for the boundary of the core in the initial model, 1 for the Interface I–II, 2 for the inner boundary of zone II, 3 for the outer boundary of zone III, and 4 for the Interface III–IV.

PHYSICAL PARAMETERS

Chemical Composition

The initial composition at age zero is assumed to be

$$X_e=0.70, \quad Y_e=0.27,$$
$$Z_e=0.03, \quad X_{\mathrm{CNO}}=\mathrm{Z}/2. \quad (1)$$

This is the same composition that Strömgren (private communication) uses for the upper main sequence. The CNO abundance was estimated from Cameron's (1959) revision of the Suess-Urey table of element abundances.

Opacity

The opacity is due to Thomson scattering by free electrons:

$$\kappa=0.19(1+X). \quad (2)$$

Equation of State

The equation of state is

$$P=(k/\mu H)\rho T+(a/3)T^4. \quad (3)$$

Energy Generation

The most recent calculation of the energy-generation rate for the CNO-cycle is due to Reeves (1962), who used the results of Caughlan and Fowler (1962). The rate may be represented with sufficient accuracy for the temperatures of our models by

$$\epsilon=\epsilon_0 X X_{\mathrm{CNO}}\rho T^{\nu}, \quad (4)$$

with $\nu=15$ and $\log\epsilon_0=-106.6$. Actually, for the early models oxygen is not quite yet in equilibrium. This formula is accurate to within 15 percent, the estimated probable error of the exact formula. The reaction $N^{14}(p, \gamma)O^{15}$ appears now to be nonresonant, so that the major uncertainty lies in the extrapolation of the experimental data on $O^{16}(p, \gamma)F^{17}$ to low energies. Fortunately, the badly known $O^{17}(p, \alpha)N^{14}$ rate does not significantly affect the energy generation for temperatures above 2×10^7 °K (Reeves 1962).

It may be noted that the values of ϵ obtained from the formulae of Fowler (1960) and Burbidge, Burbidge, Fowler, and Hoyle (1957) are the same as Reeves's values to within 15 percent, for the temperatures of our models.

BASIC EQUATIONS

The basic equilibrium equations are

$$\frac{r}{P}\frac{dP}{ar}=\frac{GM(r)\rho}{rP}=V,$$
$$\frac{r}{M(r)}\frac{dM(r)}{dr}=\frac{4\pi r^3\rho}{M(r)}=U,$$
$$\frac{T}{P}\frac{dP}{dT}=n+1, \qquad \frac{dL(r)}{dM(r)}=\epsilon, \quad (5)$$

where the radiative and adiabatic indices, n, are given by

$$(n+1)_{\mathrm{rad}}=\frac{16\pi cG(1-\beta)M(r)}{\kappa L(r)},$$
$$(n+1)_{\mathrm{ad}}=4-[1.5\beta^2/(4-3\beta)]. \quad (6)$$

In addition, we have the following two equations governing the distribution of chemical composition in zones II and III, respectively

$$[(n+1)_{\mathrm{rad}}]^{-1}=[(n+1)_{\mathrm{ad}}]^{-1}+\frac{\beta}{4-3\beta}\frac{d\log\mu}{d\log P},$$
$$\mu=\mu_\iota[M(r)/M(r)_\iota]^{-\lambda_\iota}, \quad (7)$$

where ι refers to the point at which the mass fraction is equal to that of the core of the previous

model. The composition exponents λ_i are discussed in the following section.

We introduce the following nondimensional variables:

$$r=xR, \quad M(r)=qM,$$
$$P=pP_0, \quad T=tT_0 \quad \text{(zones I–III)}, \quad (8)$$
$$r=x^*r_0, \quad M(r)=q^*M_0,$$
$$P=p^*P_c, \quad T=t^*T_c \quad \text{(zone IV)}, \quad (9)$$

where

$$P_0=GM^2/4\pi R^4, \quad T_0=\mu_e HGM/kR,$$
$$r_0=\frac{k}{\mu_c H T_c}\left(\frac{3(1-\beta_c)}{4\pi Ga}\right)^{1/2}, \quad M_0=r_0(kT_c/\mu_c HG). \quad (10)$$

Then since effectively all the nuclear energy generation takes place in the core, we may set $L(r)=L$ in zones I–III and obtain

$$dp/dx=-\beta pql/x^2t,$$
$$dq/dx=\beta px^2l/t,$$
$$dt/dx=-C(\beta pj/x^2t^4),$$
$$\beta=1-A(t^4/p) \quad \text{(zones I–III)}, \quad (11)$$

with

$$\frac{d\log l}{dx}=\left[\frac{d\log t}{dx}-\frac{1}{(n+1)_{ad}}\frac{d\log p}{dx}\right]\frac{4-3\beta}{\beta} \quad \text{(zone II)}, \quad (12)$$

and

$$l=l_i(q/q_i)^{-\lambda_i} \quad \text{(zone III)}, \quad (13)$$

where

$$l=\mu/\mu_e, \quad j=\frac{\mu}{\mu_e}\frac{1+X}{1+X_e} \qquad A=4\pi a/3(\mu_e H/k)^4 G_2 M^2,$$
$$C=0.19(1+X_e)\frac{3}{64\pi^2 ac}\left(\frac{k}{\mu_e HG}\right)^4\frac{L}{M^3}. \quad (14)$$

For the core we have

$$dp^*/dx^*=-\beta p^*q^*/x^{*2}t^*,$$
$$dq^*/dx^*=\beta p^*x^{*2}/t^*,$$
$$dt^*/dx^*=\frac{1}{(n+1)_{ad}}\frac{t^*}{p^*}\frac{dp^*}{dx^*},$$
$$\beta=1-(1-\beta_c)(t^{*4}/p^*) \quad \text{(zone IV)}. \quad (15)$$

Given μ_e and M, the solutions of zones I and II are specified by the luminosity parameter C. The solutions of zone III require an additional parameter, the composition exponent λ. The core solutions are specified by β_c.

To obtain the luminosity, we integrate the energy generation [equation (4)] over the convective core:

$$L=\epsilon_0 X_c X_{CNO}\left(\frac{\mu_e HG}{k}\right)^\nu\frac{M^{\nu+2}}{4\pi R^{\nu+3}}\left(\frac{x_4}{x_4^*}\right)^e\left(\frac{p_4}{p_4^*}\right)^2\left(\frac{t_4}{t_4^*}\right)^{\nu-2}$$
$$\times l_c^2\int_0^{x_4^*}\beta^2 p^{*2}t^{*\nu-2}x^{*2}\,dx^*. \quad (16)$$

Since L is determined directly from the parameter C, we can obtain R. Then all the physical variables in zones I–III follow from the transformations, equations (8) and the first two equations of (10). The central temperature is obtained from $T_c=T_0t_4/t_4^*$, whereupon all the physical variables in the core follow from equations (9) and the last two equations of (10).

Near the surface and center a starting series must be applied before beginning the integration of equations (11–15). We have

$$q=1, \quad p=t^4/4C, \quad t=\tfrac{1}{4}\beta_0(x^{-1}),$$
$$\beta_0=1-4AC \quad \text{(surface)} \quad (17)$$

and

$$q^*=\tfrac{1}{3}\beta_c x^{*3}, \quad p^*=1-\tfrac{1}{6}\beta_c{}^2 x^{*2},$$
$$t^*=1-\frac{1}{6}\frac{\beta_c{}^2}{(n+1)_c}x^{*2} \quad \text{(center)}, \quad (18)$$

for x close to 1 and x^* small, respectively.

CONSTRUCTION OF MODELS

The initial model contains only a radiative envelope and convective core, and is homogeneous throughout. Thereafter, an evolutionary sequence is produced by incrementing $\log C$. The method of Sakashita and Hayashi (1961) has been followed. First, families of solutions are computed for zone IV, characterized by β_c, and for zone III, characterized by λ and fitted to zone I as if there were no growing convective instability. Zone III is fitted to the core in U, V and $(n+1)$. The semiconvective zone II is now

TABLE 1.—*Evolutionary Models of a Star of 30 M⊙ During Hydrogen-burning*

	Model 0	Model 1				Model 2				Model 3				Model 4			
i	0	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
U_i	1.59	1.36	1.41	1.42	1.66	1.04	1.13	1.15	1.72	0.763	0.824	0.861	1.73	0.725	0.775	0.812	1.71
V_i	4.25	4.28	4.02	4.04	3.96	4.26	3.62	3.69	3.62	4.22	3.18	3.32	3.50	4.22	3.12	3.27	3.51
$(n+1)_i$	3.27	3.34	-------	-------	3.38	3.44	-------	-------	3.54	3.55	-------	-------	3.70	3.56	-------	-------	3.72
$\log x_i$	−0.394	−0.413	-------	-------	−0.453	−0.453	-------	-------	−0.579	−0.502	-------	-------	−0.794	−0.510	-------	-------	−0.842
$\log q_i$	−0.222	−0.203	−0.224	−0.224	−0.262	−0.178	−0.233	−0.233	−0.338	−0.152	−0.252	−0.252	−0.446	−0.148	−0.256	−0.256	−0.466
$\log p_i$	0.705	0.747	-------	-------	0.910	0.842	-------	-------	1.316	0.959	-------	-------	1.977	0.979	-------	-------	2.122
$\log t_i$	−0.530	−0.506	-------	-------	−0.458	−0.461	-------	-------	−0.323	−0.410	-------	-------	−0.123	−0.402	-------	-------	−0.082
β_i	0.844	0.823	-------	-------	0.809	0.783	-------	-------	0.740	0.735	-------	-------	0.645	0.727	-------	-------	0.627
X_i	0.700	0.700	0.699	0.693	0.583	0.700	0.691	0.667	0.373	0.700	0.664	0.611	0.112	0.700	0.657	0.600	0.069
$\log C$	−3.563	−3.500				−3.400				−3.300				−3.285			
λ	-------	1.02				1.127				1.264				1.35			
β_c	0.768	0.738				0.678				0.594				0.578			
$\log T_c$	7.562	7.573				7.594				7.635				7.649			
$\log p_c$	0.482	0.486				0.508				0.610				0.650			
$\log(L/L\odot)$	5.148	5.211				5.311				5.411				5.426			
$\log(R/R\odot)$	0.817	0.855				0.943				1.090				1.118			
$\log T_e$	4.639	4.635				4.616				4.568				4.558			
$\bar{X}$	0.700	0.633				0.527				0.420				0.405			
τ (10^6 years)	0.00	1.33				3.09				4.50				4.67			

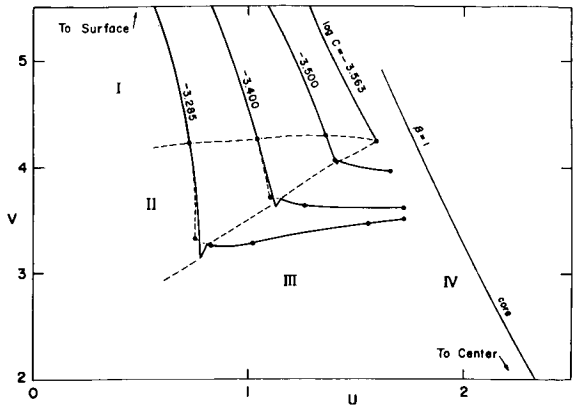


FIGURE 1.—Evolution of the model sequence in the U–V plane. The envelope solutions are designated by their log$C$ values. Dots and jumps represent the fitting points. The extended dashed lines separate the zones, which are designated by Roman numerals. The dashed lines in the solutions represent the assumed radiative zone.

taken into account by reintegrating zone I and adopting equations (12) when $(n+1)_{rad} = (n+1)_{ad}$. Since no nuclear reactions occur outside the core, the total amount of hydrogen and helium mixed must remain constant. Moreover, the convection will steepen the gradient of mean molecular weight near the Interface II–III, so that we make the approximation of specifying a discontinuity in μ here. The determination of q_2, which equals q_3, and fitting of the zones II and III are made according to the conditions

$$\int_{q_1}^{q_2} (X_e - X_{II})\, dq = \int_{q_0}^{q_3} (X_e - X_{III})\, dq,$$

$$U_3/U_2 = V_3/V_2 = \mu_3/\mu_2. \qquad (19)$$

The inward growth of zone III can be handled by a method described by Kushwaha (1957) for less massive stars. At the mass point q_t where the core of the previous model ended, zone III is integrated inward with a new trial composition exponent λ_t, using the previous set of λ's for those outer subzones of zone III not affected by mixing from the semiconvective zone. The integrations of zone III are always begun with λ_0 at q_0, the mass fraction of the core in the initial model, and are fitted to a radiative envelope (zone I) that is calculated without semiconvection.

The time step, $\Delta\tau$, is computed in the usual way (Schwarzschild, 1958, p. 196).

RESULTS AND DISCUSSION

Table 1 contains the results for the initial homogeneous model and four inhomogeneous models in which the zones were integrated with the help of an IBM 7090 computer but fitted by hand. The fitting on the $U-V$ plane is shown in figure 1, where the dots and discontinuities represent fitting points. The zones outside the core are marked off by dashed lines (not to be confused with the lines representing the assumed radiative zone that becomes semiconvective). Only the core solution for $\beta = 1$ is plotted, for the sake of clarity.

The change of mass contained in each zone during the course of the star's evolution is shown in figure 2. It is interesting to note that while the semiconvective zone grows rapidly outward, it moves inward only slightly. Figure 2 and table 1 show also that by the time the hydrogen content of the core has dropped to 10% of its initial value, both the mass fraction of the core and the mean hydrogen content of the whole star, $\bar{X}$, have decreased by about 40%.

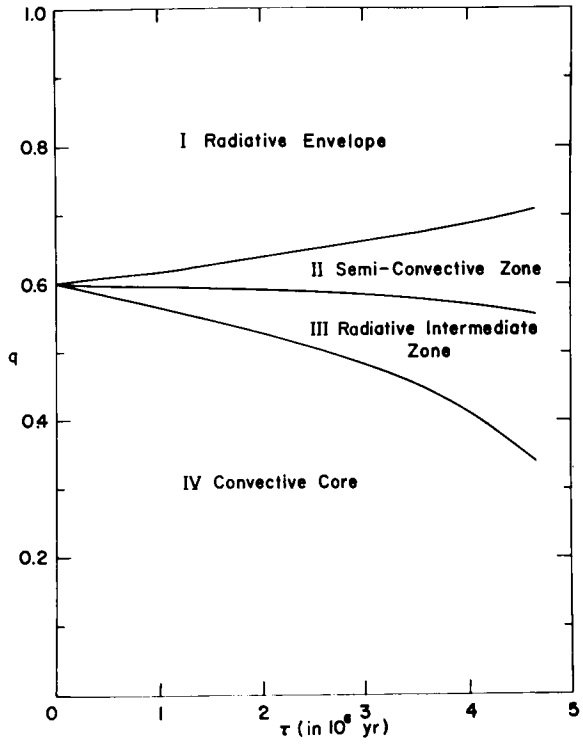


FIGURE 2.—Evolution of the structural zones in a star of 30 $M\odot$ during hydrogen-burning.

In figures 3 and 4 the hydrogen content is plotted as a function of mass fraction for the evolutionary models here considered. Dots and the vertical discontinuities represent fitting points. Clearly, the decrease of X through the semiconvective zone is far more gentle than its decrease in the radiative intermediate zone; this is to be expected from the mixing process. The jump in mean molecular weight at the interface grows to about 5% in the last model. Eventually, the discontinuity may no longer be considered a good

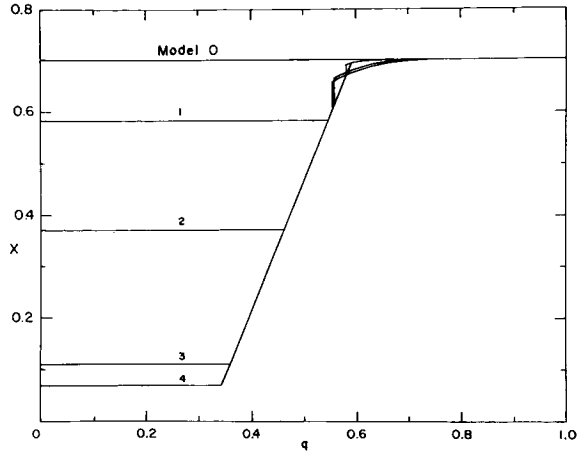


FIGURE 3.—Depletion of hydrogen as a function of mass fraction for the evolutionary models of a star of 30 $M\odot$.

approximation. However, it is clear that the semiconvective zone has little effect on the important quantities L, T_e, q_4, X_c, β_c, and T_c.

Figure 5 shows a luminosity-temperature plot of the evolutionary tracks of our model (30 $M\odot$) and the model calculated by Sakashita et al. (15.6 $M\odot$) to roughly the same amount of hydrogen depletion in the core. The more massive star has an initially steeper rise in luminosity, due to the greater mass fraction contained in the energy-producing core. When the discrepancy between the measured ultraviolet fluxes of early-type stars (Stecher and Milligan, 1962) and those predicted on the basis of model atmospheres is resolved, it will be possible to compare our results with the observations of OB clusters and associations.

Several other workers have considered the evolution of stars with mass $\approx$30 $M\odot$. In table 2 we have summarized the data on the initial model

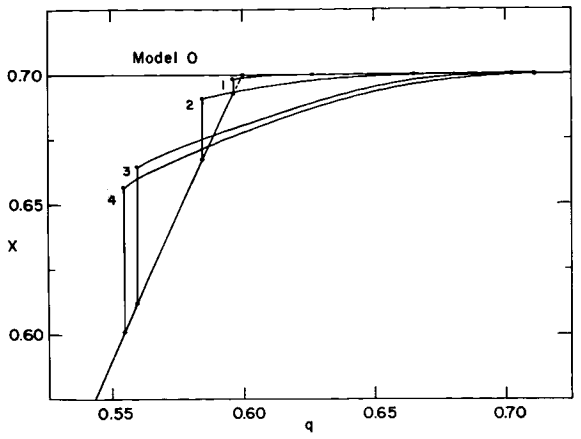


FIGURE 4.—Detail of the depletion of hydrogen in the semiconvective zone for evolutionary models of a star of 30 $M\odot$.

($X_c = X_e$) and a later model ($X_c = \frac{1}{3}X_e$), whenever the later model was given or calculable by interpolation in the given sequence. The various models are arranged in order of chemical composition (toward older population I), based primarily on the initial hydrogen content and secondarily on

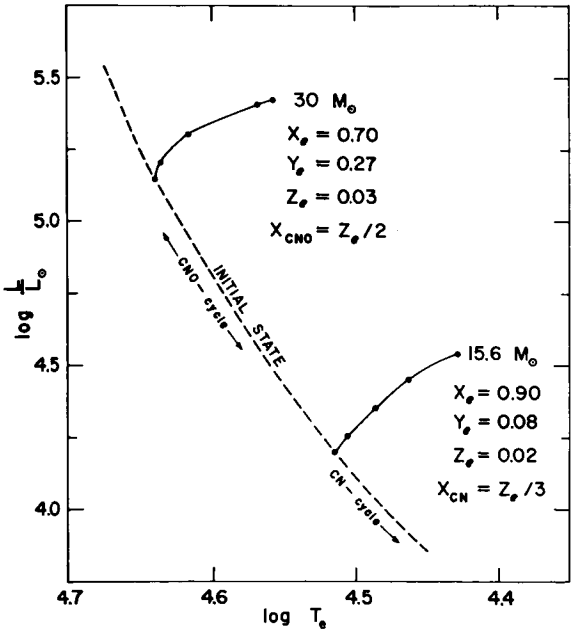


FIGURE 5.—The evolution of stars of 30 $M\odot$ (this paper) and 15.6 $M\odot$ (Sakashita, Ono, and Hayashi, 1959) is compared on a plot of luminosity vs. effective temperature.

TABLE 2.—*Comparison of Evolutionary Model Sequences for Stars With $M \approx 30\ M\odot$*

Source	$X_c = X_e$								$X_c = \frac{1}{3} X_e$		
	M ($M\odot$)	X_e	Y_c	Z_e	X_{CNO}	$\mu_e^4 M^3$	log $L/L\odot$	log T_e	log $L/L\odot$	log T_c	$\tau(10^6$ yr.)
Henyey, LeLevier, and Levee (1959)	30	0.68	0.31	0.01	0.0025	1.00	5.121	4.604	5.32	4.56	3.4
Stothers (1963)	30	0.70	0.27	0.03	0.0150	0.95	5.148	4.639	5.37	4.59	4.0
Schwarzschild and Härm (1958)	28.2	0.75	0.22	0.03	0.0043	0.68	5.035	4.630	5.30	4.57	5.2
Hoyle (1960)	30.1	0.75	0.23	0.02	0.0053	0.82	5.093	4.644	5.38	4.69	5.4
Haselgrove and Hoyle (1959)	30.2	0.75	0.24	0.01	0.0053	0.82	5.080	4.635	-------	-------	-------
Blackler (1958)	32	0.76	0.2375	0.0025	0.0025	0.94	5.018	4.621	-------	-------	-------
Blackler (1958)	32	0.85	0.1475	0.0025	0.0025	0.73	4.938	4.598	5.2	4.56	6.3

the metals content. We have by the lase equation of (14), $L \approx \mu_e^4 M^3$ for an opacity due solely to electron scattering, and have listed the quantity $\mu_e^4 M^3$, normalized to the first entry in table 2. Comparison with the trend of $\log(L/L\odot)$ is reasonably good. The small differences are due to differences in the chosen surface boundary conditions, opacity, nuclear energy generation, and method of integrating the basic equations. Moreover, because of increasing radiation pressure the mass-dependence of L is more like $M^{2.5}$ at $30\ M\odot$ (Haselgrove and Hoyle, 1959).

The arrangement of models in order of decreasing luminosity is fortuitously the same as that in order of decreasing CNO abundance, except for the rather anomalous model of Henyey et al. (1959). But since the CNO abundance regulates the rate of energy generation and varies by a factor of six in table 2, it can considerably alter the rate of consumption of hydrogen fuel. The change of time scale can better be seen by the following consideration of the total luminosity. Since $\mu_e \sim X^{-1}$ for small Z, $L \sim X^{-4}$ for approximately equal masses. Then $\tau \sim X L^{-1}$ or $\tau \sim X^5$, which gives the evident treand in the last column of table 2.

Work is currently in progress on the further evolution of a star of $30\ M\odot$ into the gravitational contraction phase during hydrogen exhaustion.

It is a pleasure to thank Dr. Hong-Yee Chiu and Dr. Leo Goldberg for advising this work. Dr. Martin Schwarzschild kindly offered some initial suggestions, and Dr. Robert Jastrow generously made available the facilities of the Institute for Space Studies. I gratefully acknowledge holding a Harvard scholarship during the academic year 1962–63.

REFERENCES

BLACKLER, J. M., *M.N.*, **118,** 37 (1958).
BURBIDGE, E. M., BURBIDGE, G. R., FOWLER, W. A., and HOYLE, F., *Rev. Mod. Phys.*, **29,** 547 (1957).
CAMERON, A. G. W., *Ap. J.*, **129,** 676 (1959).
CAUGHLAN, G. R., and FOWLER, W. A., *Ap. J.*, **136,** 453 (1962).
DEUTSCH, A. J., *Stellar Atmospheres*, ed. J. L. Greenstein (Chicago: University of Chicago Press), p. 543 (1960).
FOWLER, W. A., *Modèles d'étoiles et évolution stellaire* (extrait des *Mem. Soc. R. Sci. Liège*, 5th ser., vol. 3), p. 207 (1960).
HASELGROVE, C. B., and HOYLE, F., *M.N.*, **119,** 112 (1959).
HAYASHI, C., and CAMERON, R. C., *Ap. J.*, **136,** 166 (1962).
HENYEY, L. G., LELEVIER, R., and LEVÉE, R. D., *Ap. J.*, **129,** 2 (1959).
HOYLE, F., *M.N.*, **120,** 22 (1960).
HUANG, S.-S., MILLIGAN, J. E., and STECHER, T. P., *Pub. A.S.P.*, **74,** 312 (1962).
———, and STRUVE, O., *Stellar Atmospheres*, ed. J. L. Greenstein (Chicago: University of Chicago Press), p. 321 (1960).
KUSHWAHA, R. S., *Ap. J.*, **125,** 242 (1957).

MESTEL, L., *Ninth International Symposium, Liege*, p. 235 (1959).
MUSTEL, E. R., *I.A.U. Symposium* (Paris), **10,** 20 (1959).
REEVES, H., *Stellar Energy Sources* (Institute for Space Studies, NASA report) (1962).
SAKASHITA, S., and HAYASHI, C., *Prog. Theoret. Phys.* (Kyoto), **22,** 830 (1959).
———, *ibid.*, **26,** 942 (1961).
———, ONO, Y., and HAYASHI, C., *Prog. Theoret. Phys.* (Kyoto), **21,** 315 (1959).
SCHWARZSCHILD, M., *Structure and Evolution of the Stars* (Princeton, N.J.: Princeton University Press) (1958).
———, and HARM, R., *Ap. J.*, **128,** 348 (1958).
———, *ibid.*, **129,** 637 (1959).
STECHER, T. P., and MILLIGAN, J. E., *Ap. J.*, **136,** 1 (1962).
TAYLER, R. J., *Ap. J.*, **120,** 332 (1954).

NEUTRINO EMISSION, MASS LOSS, AND THE FREQUENCY OF SUPERNOVAE

RICHARD STOTHERS*

Goddard Space Flight Center

Without steady mass loss during the red-giant phase, most stars with mass above the Chandrasekhar limit should evolve into type II supernovae, but the predicted number of supernovae is much larger than the observed number. Although mass loss considerably lowers the predicted number, its effect will be reduced because of neutrino emission which decreases the time scale of the late phases. However, by use of presently known rates of stellar evolution and mass ejection, it is shown that loss of mass *with or without* neutrino emission will roughly account for the observed number of type II supernovae. Thus neutrino emission is not contradicted by observations of supernova frequency.

INTRODUCTION

Against the current theory of progressive nucleosynthesis in stars (Burbidge, Burbidge, Fowler, and Hoyle, 1957) is raised the serious objection that far more supernovae should be observed than is the actual case. For instance, on the assumption that every star with mass above the Chandrasekhar limit ($1.4\mathfrak{M}\odot$) gets rid of its excess mass by supernova explosion, we compute for the Galaxy with $\sim 2\times 10^9$ stars of $\mathfrak{M} > 1.4\mathfrak{M}\odot$ and average lifetime $\sim 3\times 10^9$ years (Schwarzschild, 1958) a frequency of at least 200 supernovae per 300 years. This is too large by two orders of magnitude, compared with the observed frequency of about 1 supernova every 300 years (Zwicky, 1958). This well-known result was obtained by Schwarzschild (1958) in two slightly different ways, using the death-rate function and the observed numbers of white dwarfs.

It is now apparent that many stars undergo extensive mass loss in the form of "winds" during the red-giant phase of their evolution (Deutsch, 1956a; Weymann, 1962). This is believed to account for most of the excess mass that must be expelled. On the neutrino theory of late stellar evolution, however, the time scale is so shrunken that mass loss due to winds may be ineffective, and the neutrino luminosity rises so high that energy requirements can force stars of relatively low mass to evolve to the supernova stage (see, e.g., Chiu, 1961a,b). These considerations would lead us to expect again a higher frequency of supernova outbursts. It is the purpose of this paper to see whether the theory of neutrino emission is in conflict with observations.

EVOLUTIONARY TIME SCALES

Detailed stellar models have been computed for the late evolutionary phases only in the case of three masses $\geq 1.3\mathfrak{M}\odot$ (see table 1 and references). Accordingly we list the three closest masses from the observed mass-spectrum relation tabulated by Schwarzschild (1958). The main sequence lifetime, τ_{ms}, is defined as the time required for a star to burn 13% of its hydrogen.

Hayashi and Cameron (1962a,b) and Hayashi, Hoshi, and Sugimoto (1962) have computed models for a star of $15.6\mathfrak{M}\odot$ as a red supergiant. The lifetime of carbon-burning is 2.3×10^5 years and of the later phases, $\sim 6\times 10^5$ years, without neutrino emission. These lifetimes are shortened

*Now at Harvard College Observatory.

TABLE 1.—*Stellar Lifetimes and Mass Loss During the Red-Giant Phase*

$\mathfrak{M}/\mathfrak{M}\odot$*	Sp*	$\log\tau_{ms}$† (yr)	$\log\tau_{rg}$ (yr) without ν-loss	$\log\tau_{rg}$ (yr) with ν-loss	$-\mathfrak{M}$‡ ($\mathfrak{M}\odot$/yr)	References for τ_{rg}§
18	B2	7.6	6.0	>4.9	4×10^{-6}	1,4
4	B8	8.4	>6.6	6.0	$>3\times10^{-8}$	2,4
1.4	F1	9.6	>7.4	>7.4	----------	3

*Observed mass-spectrum relation for the main sequence (Schwarzschild, 1958, p. 277).

†Computed by Schwarzschild (1958, p. 277) as the time required to burn 13% of the hydrogen.

‡Observed by Deutsch (1956) and Weymann (1962) for α Her($\sim4\,\mathfrak{M}\odot$, Weymann 1960) and α Ori($\sim20\,\mathfrak{M}\odot$), respectively.

§Sources are (1) Hayashi and Cameron (1962*a*, *b*) for a star of $15.6\,\mathfrak{M}\odot$; (2) Hayashi, Nishida, and Sugimoto (1962) for a star of $4\,\mathfrak{M}\odot$; (3) Schwarzschild and Selberg (1962) and Schwarzschild and Harm (1962) for a star of $1.3\,\mathfrak{M}\odot$; (4) improved values of Hayashi, Hoshi, and Sugimoto (1962).

to a total of at most 8×10^4 years if neutrino emission is included. Reeves (1962) also estimates that the time scale is reduced by a factor of 10 because of neutrino emission.

A star of intermediate mass ($4\mathfrak{M}\odot$) makes a brief excursion into the G and K giant region during helium-burning, but does not become a red supergiant until the onset of carbon-burning (Hayashi, Nishida, and Sugimoto, 1962; Hayashi, Hoshi, and Sugimoto, 1962). It spends 1×10^6 years burning carbon, and, extrapolating from the results on the star of $15.6\mathfrak{M}\odot$, we estimate that it spends at least 3×10^6 years in the later phases. This number may be considerably underestimated, as degeneracy in the core will occur in the case of a star of $4\mathfrak{M}\odot$, whereas in the case of $15.6\mathfrak{M}\odot$ it is very slight. In fact, Hayashi, Nishida, and Sugimoto (1963) found that $\sim4\mathfrak{M}\odot$ represents a critical mass below which degeneracy occurs in the core just before helium-burning. The red-giant lifetime with neutrino emission included is 10^6 years (Hayashi, Hoshi, and Sugimoto, 1962).

The evolutionary track of a star of $1.3\mathfrak{M}\odot$ enters the G–K red-giant region during gravitational contraction and the onset of helium-burning. Degeneracy in the core raises the time scale to $\tau_{rg}>2.5\times10^7$ years (Schwarzschild and Selberg, 1962; Schwarzschild and Härm, 1962). Presumably, after a brief period of nondegeneracy following the helium flash, the further evolution of the star is finally limited by renewed degeneracy, and the star becomes a white dwarf or type I supernova (population II and due to an entirely different mechanism; see Hoyle and Fowler, 1960) without returning to the red-giant region. Plasma neutrino emission will have some effect on the luminosity of the core of the star, but not on the total luminosity (Chiu, 1963b), and is therefore not believed to affect the time scale significantly.

Although the models for these three stars had different assumed initial chemical compositions, the time scales of evolution, especially during the later phases, should not be affected too much by a change in initial composition, at least to within the accuracy we require (cf. table 1).

MASS LOSS

We now turn to the problem of mass loss by stars in the form of winds or ejection of shells. The solar wind carries away $3\times10^{-14}\mathfrak{M}\odot$/year, which is completely negligible (Parker, 1963). Underhill (private communication) obtains observationally a mass loss of 10^{-6} to $10^{-7}\mathfrak{M}\odot$/year for Wolf-Rayet stars and believes that the normal O stars would have this rate reduced by a factor of 10. For a Be shell star ($10\mathfrak{M}\odot$), Underhill (1960) obtains $10^{-7}\mathfrak{M}\odot$/year, to be reduced by a similar factor for the normal B stars. Hence with main-sequence lifetimes of 5×10^6 and 1×10^8 years, respectively, the mass loss even from O and B stars is negligible during the main-sequence phase of evolution.

During the red-giant phase, however, obser-

vations show that mass loss can be considerable. Weymann (1962) obtains for α Orionis an outflow of $4\times10^{-6}\mathfrak{M}\odot$/year. Allen (1955) gives for a star with its spectral type (M2 Iab) a mass of $\sim$20 $\mathfrak{M}\odot$. The corresponding escape velocity is consistent with the assumption of outflow of material. Deutsch (1956a) obtains a mass loss of 3×10^{-8} $\mathfrak{M}\odot$/year for α Herculis. This star is a visual binary with a computed mass of $15\mathfrak{M}\odot$ for the more massive M component. However, its spectral and luminosity class (M5 II) would indicate a mass more like $4\mathfrak{M}\odot$ (Allen 1955). Moreover, Weymann (1960) has pointed out that the interpreted velocity of outflow is not consistent with the escape velocity calculated for $15\mathfrak{M}\odot$. He uses Wilson's (1960) distance determination for α Herculis and inspection of the H–R diagram of galactic clusters to suggest that the mass of the M component is $\sim4\mathfrak{M}\odot$. We adopt this mass in table 1.

Hayashi, Hoshi, and Sugimoto (1962) found that the main effect of mass loss is to reduce the effective temperature. Now Deutsch (1960) showed that mass loss is vastly more efficient at the lower temperatures. Since stars with $\mathfrak{M}\geq4$ $\mathfrak{M}\odot$ spend less time in carbon-burning than in the later phases (without neutrino emission), the average rate of mass loss will be greater than that indicated in table 1. Moreover, Deutsch's value of the loss from α Herculis is a lower limit, and he points out that significant loss may even occur at the earlier spectral types, which would include part of the helium-burning phase of a star of $4\mathfrak{M}\odot$ (Deutsch, 1960). Since the rate of loss should be roughly independent of stellar mass, perhaps Weymann's value is a better average. Therefore it seems that mass loss will be effective in reducing the mass of a star with initial $4\mathfrak{M}\odot$ to below the Chandrasekhar limit. The star will then become either a white dwarf with a core composed of the products of carbon-burning or perhaps a supernova of Type I. Table 1 also suggests that mass loss will be effective even up to 15–$20\mathfrak{M}\odot$ (main-sequence spectrum B2) for the case of no neutrino loss. Stars with $\mathfrak{M}<4\mathfrak{M}\odot$ will almost certainly lose some mass, cool off, and avoid type II supernova explosion, with or without neutrino emission.

If mass loss is not efficient enough, however, ordinary evolutionary processes will make supernovae out of all stars with $\mathfrak{M}>4\mathfrak{M}\odot$ (B8); with neutrino emission, mass loss will certainly be inefficient for the more massive stars because of the drastically reduced time scale, as indicated in table 1 for $18\mathfrak{M}\odot$ (B2). So we shall regard $\sim$B8 as the critical spectral class for which earlier stars become supernovae on the neutrino theory.

We note, parenthetically, that observations of globular and open clusters support the theoretical results that main-sequence stars evolve into red giants (at least the stars of low and intermediate mass). For the massive stars the evidence from OB clusters is suggestive (Hayashi and Cameron, 1962a,b), even though not required (Chiu, 1963a). Moreover, star-counts and the distribution of M giants in the field (Deutsch, 1956b) support this view of evolution.

STELLAR STATISTICS AND FREQUENCY OF SUPERNOVAE

We must now estimate the number of O and B stars in the Galaxy. Roberts (1957) used Shmirelman's (1952) value of 1.8×10^6 B0–B5 stars to obtain 2.5×10^6 B0–B7 stars. Rubin, Burley, Kiasatpoor, Klock, Pease, Rutscheidt, and Smith (1962) list 1440 O–B5 stars within a radius of 3 kpc around the sun; this value is estimated to be 5 per cent complete. If we assume with Rubin *et al.* that the radius of the Galaxy is 15 kpc, we obtain a total of 7×10^5 O–B5 stars in the Galaxy. By extrapolation we get 1×10^6 O–B7 stars. Parenago (1948) and Roberts (1957) obtain 6×10^3 O stars in the Galaxy. Therefore we estimate that there are 5×10^4 O–B2 stars. Using the main-sequence lifetime of the latest star in a small spectral

TABLE 2.—*Predicted Frequency of Supernovae for Various Spectral Groups*

Sp. group	No. of stars	log τ_{ms} yr	SN/300 yr
O	6×10^3	6.6	0.5
O–B2	5×10^4	7.6	1
O–B5	7–18×10^5	8.0	4–6
O–B7	1–2.5×10^6	8.3	6–10

interval, we divide the number of stars in the interval by τ_{ms} to obtain the predicted number of supernovae (SN). The cumulative spectral intervals and numbers of supernovae are listed in table 2.

The observed frequency of both type I and type II supernovae in external galaxies is one in 300–400 years (Zwicky, 1958). The uncertain completeness of the searches and intragalactic extinction indicate a true frequency that is higher by an unknown factor. The three, well-authenticated supernovae in our own Galaxy were of type I and show an apparent frequency of 1 SN/300 years. Other supernovae have been recorded in our Galaxy (Kukarkin, Parenago, Efremov, and Kholopov, 1958), but their type is unknown. However, type II supernovae are known to be more common in general (Payne-Gaposchkin, 1957). Hence we shall accept for our own Galaxy the "true" extragalactic rate of type II supernovae as >1 SN/300 years.

We recall from the first paragraph of this paper that without mass loss the frequency of supernovae should be greater than 200 per 300 years. This is too large by about two orders of magnitude. From table 2 we note that on the basis of mass loss *with or without* neutrino emission the predicted frequencies (despite the uncertainties in arriving at them) fall within a reasonable range of the observed frequency. In fact, if Payne-Gaposchkin's (1957) estimate of at least 8 type II SN/300 years in the Galaxy is correct, the neutrino hypothesis improves the agreement.

Hence we conclude that the assumption of neutrino emission is at least not contradicted by observations of supernova frequency.

I am indebted to Drs. Hong-Yee Chiu and A. G. W. Cameron for criticisms and discussion. It is a pleasure to thank Dr. Robert Jastrow for the hospitality of the Institute for Space Studies.

REFERENCES

ALLEN, C. W., *Astrophysical Quantities* (London: Athlone Press) (1955).

BURBIDGE, E. M., BURBIDGE, G. R., FOWLER, W. A., and HOYLE, F., *Revs. Mod. Phys.*, **29**, 547 (1957).

CHIU, H.-Y., *Ann. Phys.*, **15**, 1 (1961*a*).

———, *ibid.*, **16**, 321 (1961*b*).

———, *A.J.*, **68**, 70 (1963*a*).

———, *Ap. J.*, **137**, 343 (1963*b*).

DEUTSCH, A. J., *Ap. J.*, **123**, 210 (1956*a*).

———, *Pub. A.S.P.*, **68**, 308 (1956*b*).

———, *Stellar Atmospheres*, ed. J. Greenstein (Chicago: University of Chicago Press), p. 543 (1960).

HAYASHI, C., and CAMERON, R. C., *Ap. J.*, **136**, 166 (1962*a*).

———, *A.J.*, **67**, 577 (1962*b*).

———, HOSHI, R., and SUGIMOTO, D., *Prog. Theoret. Phys. Suppl.* (Kyoto), No. 22 (1962).

———, NISHIDA, M., and SUGIMOTO, D., *Prog. Theoret. Phys.* (Kyoto), **27**, 1233 (1962).

HOYLE, F., and FOWLER, W. A., *Ap. J.*, **132**, 565 (1960).

KUKARKIN, B. V., PARENAGO, P. P., EFREMOV, Y. I., and KHOLOPOV, P. N., *General Catalogue of Variable Stars* (Moscow: Academy of Sciences of the U.S.S.R. Press) (1958).

PARENAGO, P. P., *Russian A.J.*, **25**, 123 (1948).

PARKER, E. N., Colloquium, Institute for Space Studies, New York (1963).

PAYNE-GAPOSCHKIN, C., *The Galactic Novae* (New York: Interscience Publishers) (1957).

REEVES, H., *Ap. J.*, **138**, 79 (1963).

ROBERTS, M. S., *Pub. A.S.P.*, **69**, 59 (1957).

RUBIN, V. C., BURLEY, J., KIASATPOOR, A., KLOCK, B., PEASE, G., RUTSCHEIDT, E., and SMITH, C., *A.J.*, **67**, 491 (1962).

SCHWARZSCHILD, M., *Structure and Evolution of the Stars* (Princeton, N.J.: Princeton University Press) (1958).

———, and HARM, R., *Ap. J.*, **136**, 158 (1962).

———, and SELBERG, H., *Ap. J.*, **136**, 150 (1962).

SHNIRELMAN, P. G., *Russian A.J.*, **29**, 179 (1952).

UNDERHILL, A. B., *Stellar Atmospheres*, ed. J. Greenstein (Chicago: University of Chicago Press), p. 411 (1960).

WEYMANN, R., *Ap. J.*, **132**, 380 (1960).

———, *ibid.*, **136**, 844 (1962).

WILSON, O. C., *Ap. J.*, **131**, 75 (1960).

ZWICKY, F., *Handbuch der Physik*, ed. S. Flugge (Berlin: Springer-Verlag), **51**, 766 (1958).

II. CELESTIAL MECHANICS AND GEODESY

A CONTRIBUTION TO THE THEORY OF CRITICAL INCLINATION OF CLOSE EARTH SATELLITES

SHINKO AOKI*

Goddard Space Flight Center

A higher order theory of the motion of close satellites in the vicinity of the critical inclination is developed for the case of small eccentricity as well as moderate eccentricity. Terms up to the third order of magnitude are included for a potential which is plane symmetrical. It is shown that solutions up to this order can be obtained by using only Jacobian elliptic functions for the case of small eccentricity. However, for moderate eccentricity, elliptic integrals of the second kind or the Jacobian zeta-function are required. To obtain further approximations, elliptic integrals of the third kind are necessary for the case of small eccentricity; but, for the case of moderate eccentricity only the Jacobian elliptic functions and elliptic integrals of the second kind are required. In any case, the solutions are, in general, periodic functions—as is well known from the form of the equations of motion.

INTRODUCTION

Many authors have shown their interest in teh motion of close earth satellites near the critical inclination (for example, references 1–5). However, almost all of these authors have considered only the first terms, which are of the order of the square of J_2 in the equations of motion. However, it was first shown by Izsak[5] that continuations of the ordinary treatment of the libration to the higher approximations would break down; and, in the case of small eccentricity, some libration of a peculiar kind including the next order of magnitude would occur. The first term, which has e_0^2 as a factor, is so small that the next order term—which would be considered as a higher order of magnitude in the normal case—becomes the same order of magnitude as the preceding term. However, he presented this only for a consideration of the form of the Hamiltonian; he did not try to solve the equations of motion with time as the independent variable.

Here some expressions for these solutions will be presented, not only for the case of small eccentricity, but also for the normal case including terms up to the third order of J_2 in the original equations of motion, where J_2 is the coefficient of second zonal harmonics of the earth's potential and J_4 is assumed to be of the order of the square of J_2 and, further, J_6 to be of the third order of magnitude. The odd harmonics will be totally neglected. A higher order theory including these terms is under consideration, but must be treated in the future.

The method adopted is to introduce intermediary solutions, of which the argument λu is only slightly different from time (the independent variable) except for a constant factor, and such that by which the amplitude of the libration or the change of states—in other words, a transition from revolution to libration—is fully determined. In the case of small eccentricity, the intermediary solution can be obtained by the following integration:

$$\frac{d\omega_1}{\sqrt{(1-\alpha\sin^2\omega_1)(1-\beta\sin^2\omega_1)}} = \lambda\, du\ ,$$

*NAS–NASA Research Associate.

where $\pi/2-\omega_1$ is the mean value of the argument of perigee and λ is constant; this integration can be performed even if only the Jacobian elliptic functions are used, as will be seen in a later section.*

Further approximations, beyond the aforementioned order of magnitude, would require more complicated calculations. Some comments will be made on the subject although explicit formulas are not included. Izsak, after removing the so-called periodic perturbations, wrote the equations of motion:

$$\frac{d\rho}{d\tau} = \frac{\partial \Phi}{\partial \omega} ,$$

$$\frac{d\omega}{d\tau} = -\frac{\partial \Phi}{\partial \rho} ,$$

where ω is the mean value of the argument of perigee; ρ is some function of G, the conjugate variable to ω; and

$$\Phi = A(\rho) + j B(\rho) \cos^2 \omega ,$$

$$A(\rho) = \rho^2 + \sum_3 \alpha_n \rho^n ,$$

$$B(\rho) = \sum_0 \beta_n \rho^n ,$$

$$\beta_0 = e_0^2 \left[1 + O(j)\right] ,$$

$$\beta_1 = -2\left[1 + O(j)\right] ,$$

$$j = -\frac{1}{5}\left(\frac{J_4}{J_2} + J_2\right) p_0^{-2} > 0 ,$$

$$\tau = \frac{3}{4}\frac{J_2 n_0}{\zeta p_0^2}(t - t_0) .$$

Since τ has a factor of J_2, the terms of the order of j in Φ correspond to those of the order of J_2^2 or J_4 in the original equations of motion.

*A list of symbols is given in app. A.

In this paper we have started with a Hamiltonian similar to the one used by Izsak but have added some additional terms in order to more easily consider the higher order terms; although this explicit evaluation, from the coefficients of zonal harmonics of the earth's potential and from the integration constants, is not given. We are concerned with the solutions of the equations of motion which contain the Hamiltonian up to the order of j^2:

$$\Phi = A(\rho) + j B(\rho) \cos^2 \omega + j^2 C(\rho) \cos^4 \omega ,$$

although some comments will be made on obtaining higher order approximations.

EQUATIONS OF MOTION

We shall now consider the following equations:

$$\left.\begin{aligned} \frac{d\rho}{d\tau} &= \frac{\partial \Phi}{\partial \omega} , \\ \frac{d\omega}{d\tau} &= -\frac{\partial \Phi}{\partial \rho} , \end{aligned}\right\} \qquad (1)$$

where

$$\Phi = A(\rho) + j B(\rho) \cos^2 \omega + j^2 C(\rho) \cos^4 \omega ,$$

$$A(\rho) = \sum_{n=2}^{\infty} \alpha_n \rho^n ,$$

$$B(\rho) = \sum_0^{\infty} \beta_n \rho^n ,$$

$$C(\rho) = \sum_0^{\infty} \gamma_n \rho^n ,$$

and α_n, β_n, γ_n are constants which satisfy $\alpha_2>0$, $\beta_0>0$.

In order to remove the first degree term of ρ in the Hamiltonian function, we change the variable ρ to $\chi=\rho-\varrho$ $(\cos^2\omega)$ such that

$$\left(\frac{\partial \Phi}{\partial \chi}\right)_{\chi=0} = \left(\frac{\partial \Phi}{\partial \rho}\right)_{\rho=\varrho} = 0 . \qquad (2)$$

If we take the terms up to j^2 in equation 2, we

have

$$j\beta_1 \cos^2\omega + j^2\gamma_1 \cos^4\omega + 2\left(\alpha_2 + j\beta_2 \cos^2\omega\right)\rho + 3\alpha_3\rho^2 = 0 ,$$

from which

$$\rho = -\frac{1}{2} j \frac{\beta_1}{\alpha_2} \cos^2\omega + j^2 \left(-\frac{1}{2}\frac{\gamma_1}{\alpha_2} + \frac{1}{2}\frac{\beta_1\beta_2}{\alpha_2^2} - \frac{3}{8}\frac{\alpha_3\beta_1^2}{\alpha_2^3}\right)\cos^4\omega . \tag{3}$$

Thus, the equations,

$$\frac{d\chi}{d\tau} = \frac{\partial\Psi}{\partial\omega} ,$$

$$\frac{d\omega}{d\tau} = -\frac{\partial\Psi}{\partial\chi} , \tag{4}$$

are obtained with

$$\Psi(\chi, \omega) = \Phi(\rho + \chi, \omega)$$

$$= a_0 + a_2\chi^2 + a_3\chi^3 + a_4\chi^4 + \cdots ,$$

where

$$a_0 = j\beta_0 \cos^2\omega + j^2\left(-\frac{1}{4}\beta_1^2\alpha_2^{-1} + \gamma_0\right)\cos^4\omega + \cdots ,$$

$$a_2 = \alpha_2 + j\left(\beta_2 - \frac{3}{2}\alpha_3\beta_1\alpha_2^{-1}\right)\cos^2\omega + \cdots , \tag{5}$$

$$a_3 = \alpha_3 + \cdots ,$$

$$a_4 = \alpha_4 + \cdots .$$

We have considered terms up to the order of j^2 in Ψ, assuming χ is the order of $j^{1/2}$. We know that

$$\Psi = j\Gamma , \tag{6}$$

where Γ is constant, is an integral of equation 4. Neglecting higher order terms, we can easily obtain an expression for χ from equation 6:

$$\chi = \chi_1 - \frac{1}{2}a_3 a_2^{-1}\chi_1^2 + \left(-\frac{1}{2}a_4 a_2^{-1} + \frac{5}{8}a_3^2 a_2^{-2}\right)\chi_1^3 , \tag{7}$$

where

$$\chi_1^2 = \left(j\Gamma - a_0\right)a_2^{-1} ;$$

$$= \alpha_2^{-1} j\left[\Gamma - \beta_0\cos^2\omega + j\left(\frac{1}{4}\beta_1^2\alpha_2^{-1} - \gamma_0\right)\cos^4\omega \right.$$

$$\left. + j\left(\Gamma - \beta_0\cos^2\omega\right)\left(-\beta_2\alpha_2^{-1} + \frac{3}{2}\alpha_3\beta_1\alpha_2^{-2}\right)\cos^2\omega\right] .$$

Rewriting χ_1^2 in the form

$$\chi_1^2 = \mu\left(1 + P_1\cos^2\omega + Q_1\cos^4\omega\right) ,$$

where

$$\mu = \Gamma a_2^{-1} j ,$$

$$P_1 = \Gamma^{-1}\left[-\beta_0 + j\Gamma\left(-\beta_2 a^{-1} + \frac{3}{2} a_3 \beta_1 a_2^{-2}\right)\right] ,$$

$$Q_1 = j\Gamma^{-1}\left(\frac{1}{4}\beta_1^2 a_2^{-1} - \gamma_0 + \beta_0\beta_2 a_2^{-1} - \frac{3}{2} a_3 \beta_0 \beta_1 a_2^{-2}\right) ,$$

and putting $\omega_1 = (\pi/2) - \omega$, we have

$$\frac{d\omega_1}{2a_2 x + 3a_3 x^2 + 4a_4 x^3} = \frac{d\omega_1}{2a_2 \mu^{1/2} x_0 \left[1 + a_3 a_2^{-1} \mu^{1/2} x_0 + \left(\frac{3}{2} a_4 a_2^{-1} - \frac{7}{8} a_3^2 a_2^{-2}\right)\mu x_0^2\right]} \qquad (8)$$

$$= d\tau ,$$

where

$$\left.\begin{aligned} x_0 &= a_2 a_2^{-1} \mu^{-1/2} x_1 \\ &= \sqrt{1 + P\sin^2\omega_1 + Q\sin^4\omega_1} \\ P &= -\beta_0\Gamma^{-1} + j\left(\beta_2 a_2^{-1} - \frac{3}{2} a_3 \beta_1 a_2^{-2}\right) \\ Q &= j\Gamma^{-1}\left(\frac{1}{4}\beta_1^2 a_2^{-1} - \gamma_0 - \beta_0\beta_2 a_2^{-1} + \frac{3}{2} a_3 \beta_0 \beta_1 a_2^{-2}\right) . \end{aligned}\right\} \qquad (9)$$

INTERMEDIARY SOLUTIONS

From the denominator on the right side of equation 8 we can see that it will be zero if, and *only* if, $x_0 = 0$, provided $\mu = 0(j\Gamma)$ is small. This means that the essential character of the trajectory is determined entirely by x_0. In the case ω_1 is real—which includes, in general, j^2 terms of the equation of motion. Thus, x_0 discriminates among three kinds of motion: libration, revolution, and the critical case.

Therefore, we first must consider

$$\frac{d\omega_1}{\sqrt{1 + P\sin^2\omega_1 + Q\sin^4\omega_1}} = \frac{d\omega_1}{\sqrt{(1 - \alpha\sin^2\omega_1)(1 - \beta\sin\omega_1)}} \qquad (10)$$

$$= du .$$

Putting $\sin^2\omega_1 = x$, we have

$$du = \frac{dx}{2\sqrt{(1-x)\,x\,(1 + Px + Qx^2)}} ; \qquad (11)$$

and, putting $y=x^{-1}$, we have

$$du = \frac{-dy}{\sqrt{4(y-1)(y^2+Py+Q)}} = \frac{-dy}{\sqrt{4(y-1)(y-\alpha)(y-\beta)}},$$

or

$$du = \frac{-dz}{\sqrt{4(z-e_f)(z-e_g)(z-e_h)}},$$

where

$$z = y + \frac{P-1}{3},$$

and

$$(e_f, e_g, e_h) = \left(\frac{P}{3} + \frac{2}{3}, \alpha + \frac{P-1}{3}, \beta + \frac{P-1}{3}\right).$$

Thus, we have

$$z=\wp(u-u_0),$$

where $\wp u$ is the Weierstrassian elliptic function of the second order. We assume that the integration constant is such that

$$u_0 = 0,$$

which corresponds to the initial condition of $\omega_1=0$ when $u=0$, as will be seen later. Then we have

$$\sin^2\omega_1 = \left(\wp u - \frac{P-1}{3}\right)^{-1} = (\wp u - e_f + 1)^{-1}$$

$$= (1 + fj^2 u)^{-1},$$

or

$$\cot\omega_1 = \pm fju,$$

and

$$\sqrt{1 + P\sin^2\omega_1 + Q\sin^4\omega_1} = gju \cdot hju(1 + fj^2 u)^{-1}.$$

In the above expressions, fju is an associated primitive elliptic function of the first order,[6] whose poles are $u=0$ and $2\omega_g$ and whose zeros are $u=\omega_f$ and $\omega_f+2\omega_g$ within a parallelogram $(2\omega_f, 4\omega_g)$, where we have

$$\omega_p = \int_{e_p}^{\infty} \frac{dz}{\sqrt{4(z-e_f)(z-e_g)(z-e_h)}},$$

with $p=f, g, h$ and $\omega_f+\omega_g+\omega_h=0$. The other gju and hju are given by permutations of (f, g, h).

However, this expression is so general that, at first glance, we cannot see how the function behaves for real values of u, which is the only case of interest. Therefore, in order to clarify its behavior, we shall divide the problem into several cases.

(*i*) α, β *real*: *Libration and critical cases*

(*a*) $1\leq\beta<\alpha\leq y$, *libration within* $-\alpha^{-1/2}\leq\sin\omega_1\leq\alpha^{-1/2}$

Since

$$\wp u = 1 + \frac{P-1}{3} + \frac{\alpha - 1}{\mathrm{sn}^2(\lambda u, k)}, \tag{12}$$

where $\lambda^2=\alpha-1$ and $k^2=(\beta-1)/(\alpha-1)$, we have

$$\sin^2\omega_1 = \frac{\mathrm{sn}^2(\lambda u, k)}{\alpha - \mathrm{cn}^2(\lambda u, k)} \tag{13}$$

$$\chi_0 = \sqrt{1 + P\sin^2\omega_1 + Q\sin^4\omega_1} = \frac{(\alpha-1)\,\mathrm{cn}\,\lambda u\,\mathrm{dn}\,\lambda u}{\alpha - \mathrm{cn}^2(\lambda u, k)} .$$

The case $1\leq y\leq\beta\leq\alpha$ does not correspond to any real case in the theory of the critical inclination of close artificial satellites[5]; therefore, we omit this case here, although similar expressions are obtainable if we substitute $\sin\omega=\cos\omega_1$ instead of $\sin\omega_1$ in equation 1.

(b) $\beta<1<\alpha\leq y$, *libration within* $-\alpha^{-1/2}\leq\sin\omega_1\leq\alpha^{-1/2}$

Since

$$\wp u = \beta + \frac{P-1}{3} + \frac{\alpha-\beta}{\mathrm{sn}^2(\lambda u, k)} ,$$

where $\lambda^2=\alpha-\beta$ and $k^2=(1-\beta)/(\alpha-\beta)$, we have

$$\sin^2\omega_1 = \frac{\mathrm{sn}^2(\lambda u, k)}{\alpha - \beta\,\mathrm{cn}^2(\lambda u, k)}$$

and

$$\chi_0 = \frac{(\alpha-\beta)\,\mathrm{cn}(\lambda u, k)}{\alpha - \beta\,\mathrm{cn}^2(\lambda u, k)} .$$

(c) $\beta\leq\alpha<1\leq y$, *revolution*

Since

$$\wp u = \beta + \frac{P-1}{3} + \frac{1-\beta}{\mathrm{sn}^2(\lambda u, k)} ,$$

where $\lambda^2=1-\beta$ and $k^2=(\alpha-\beta)/(1-\beta)$, we have

$$\sin^2\omega_1 = \frac{\mathrm{sn}^2(\lambda u, k)}{1 - \beta\,\mathrm{cn}^2(\lambda u, k)} ,$$

and

$$\chi_0 = \frac{(1-\beta)\,\mathrm{dn}(\lambda u, k)}{1 - \beta\,\mathrm{cn}^2(\lambda u, k)} .$$

(a′) $1<\beta=\alpha\leq y$, *critical*

Since $sn\lambda u\rightarrow\tanh\lambda U$ when $k\rightarrow1$, we have

$$\sin^2\omega_1 = \frac{\sinh^2\lambda u}{\alpha\cosh^2\lambda u - 1} ,$$

where $\lambda^2=\alpha-1$ and

$$\chi_0 = \frac{\alpha-1}{\alpha\cosh^2\lambda u - 1} .$$

(b′) $\beta<1=\alpha\leq y$, *critical*

From either case *b* or *c*, when $k\rightarrow1$, we have

$$\sin^2\omega_1 = \frac{\sinh^2\lambda u}{\cosh^2\lambda u - \beta}$$

and

$$\chi_0 = \frac{(1-\beta)\cosh\lambda u}{\cosh^2\lambda u - \beta} ,$$

where $\lambda^2=1-\beta$.

(ii) α, β *imaginary*: *Revolution*

Here the direct transformation from $\wp$-functions to Jacobian elliptic functions does not give a real function. Therefore, another type of transformation is used, as is described by Whittaker and Watson (reference 7, p. 513), since P and Q are real.

If we apply the transformation

$$y = C\frac{x-B}{x-A} ,$$

where

$$A = \frac{-1-\sqrt{1+P+Q}}{P+Q} ,$$

$$B = \frac{-1 + \sqrt{1+P+Q}}{P+Q},$$

$$C = \frac{-2 - P - Q - 2\sqrt{1+P+Q}}{P+Q},$$

to equation 11, we then have

$$du = \frac{d\omega_1}{\sqrt{1 + P\sin^2\omega_1 + Q\sin^4\omega_1}} = \frac{dy}{2\lambda\sqrt{(1-y^2)(k'^2+k^2y^2)}},$$

where

$$\lambda^2 = \sqrt{1+P+Q} = \sqrt{(1-\alpha)(1-\beta)},$$

$$k^2 = \frac{-P-2+2\sqrt{1+P+Q}}{4\sqrt{1+P+Q}} = \frac{\alpha+\beta-2}{4\sqrt{(1-\alpha)(1-\beta)}} + \frac{1}{2}.$$

Accordingly, we obtain

$$y = \mathrm{cn}\left[2\lambda(u-u_0), k\right]. \qquad (14)$$

It should be noted that the form of y in equation 14 yields a real function, when $\lambda^2 \geq 0$ and $1 \geq k^2 \geq 0$. This is made possible by choosing an appropriate sign of the above root function (which is, of course, a double-valued function) in such a way that

$$\sqrt{1+P+Q} \geq 0;$$

this is possible because P and Q are real and satisfy the inequalities

$$1 + P + Q = \left(1 + \frac{P}{2}\right)^2 - \frac{1}{4}(P^2 - 4Q) \geq 0,$$

$$(P+2)^2 \leq 4(1+P+Q),$$

since

$$\Delta = P^2 - 4Q \leq 0.$$

Choosing the integration constant $u_0 = 0$, and using the addition theorem, we have the following expressions:

$$x = \sin^2\omega_1 = \frac{\mathrm{sn}^2(\lambda u, k)}{1 - \mathrm{cn}^2\lambda u + \lambda^2\,\mathrm{cd}^2\lambda u}$$

and

$$\chi_0 = \frac{\lambda^2(1 - k^2\,\mathrm{sn}^2\lambda u\,\mathrm{cd}^2\lambda u)}{1 - \mathrm{cn}^2\lambda u + \lambda^2\,\mathrm{cd}^2\lambda u}.$$

Consider the critical case when the descriminant Δ goes to zero. If $-P = \alpha+\beta > 2$, then k^2 approaches to unity; and the above expressions, of course, tend to case *ia'*. On the other hand, if $-P = \alpha+\beta < 2$, then $k^2 \to 0$; and this case tends to be a special case of case *ic*.

(*iii*) $\alpha = \beta = 1$, *critical.*

In this case Equation 10 may be expressed directly as

$$d\omega/\sin^2\omega = -du,$$

from which we obtain $\sin\omega = 1/(u^2+1)^{1/2}$, or

$$\sin\omega_1 = u/(u^2+1)^{1/2},$$

where the integration constant is chosen such that $\omega_1 = 0$ when $u = 0$. Consequently,

$$\chi_0 = \cos^2\omega_1 = (u^2+1)^{-1}.$$

Table 1 summarizes the data from the three cases.

THE CASE OF SMALL ECCENTRICITY

If $\beta_0 (\approx e_0^2)$ and Γ are of the same order of magnitude as j, then the argument given for the equations of motion becomes somewhat simpler because χ is of the order of j so that the terms $\alpha_3\chi^3$ and $\alpha_4\chi^4$ in equations 5 are negligible provided we include terms up to the second order in these equations. Moreover, in the expression $\mu\chi_0^2$ it is sufficient to consider only terms up to the second order. In the case of small eccentricity near the critical inclination, γ_0 itself (equations 5) would be of the order of j^2, which is also negligible.[8]

Thus, in this case, the intermediary solutions described in a previous section would give sufficient accuracy provided that we take Equations 1 to the second order.

TABLE 1.—*List of Solutions for* $d\omega_1 = \chi_0\, du$, $\chi_0^2 = (1-\alpha x)(1-\beta x)$, $x = \sin^2\omega_1 = \cos^2\omega$

Case	Designation	x	χ_0	k^2	λ^2	Character of the Motion
ia	$1 \le \beta \le \alpha \le x^{-1}$	$\dfrac{\mathrm{sn}^2(\lambda u, k)}{\alpha - \mathrm{cn}^2(\lambda u, k)}$	$\dfrac{(\alpha-1)\,\mathrm{cn}\,\lambda u\,\mathrm{dn}\,\lambda u}{\alpha - \mathrm{cn}^2(\lambda u, k)}$	$\dfrac{\beta-1}{\alpha-1}$	$\alpha - 1$	libration between $-\alpha^{-1/2} \le \sin\omega_1 \le \alpha^{-1/2}$
ib	$\beta \le 1 < \alpha \le x^{-1}$	$\dfrac{\mathrm{sn}^2(\lambda u, k)}{\alpha - \beta\,\mathrm{cn}^2(\lambda u, k)}$	$\dfrac{(\alpha-\beta)\,\mathrm{cn}\,\lambda u}{\alpha - \beta\,\mathrm{cn}^2(\lambda u, k)}$	$\dfrac{1-\beta}{\alpha-\beta}$	$\alpha - \beta$	libration between $-\alpha^{-1/2} \le \sin\omega_1 \le \alpha^{-1/2}$
ic	$\beta \le \alpha < 1 < x^{-1}$	$\dfrac{\mathrm{sn}^2(\lambda u, k)}{1 - \beta\,\mathrm{cn}^2(\lambda u, k)}$	$\dfrac{(1-\beta)\,\mathrm{dn}\,\lambda u}{1 - \beta\,\mathrm{cn}^2(\lambda u, k)}$	$\dfrac{\alpha-\beta}{1-\beta}$	$1 - \beta$	revolution
ia'	$1 < \beta = \alpha < x^{-1}$	$\dfrac{\sinh^2 \lambda u}{\alpha \cosh^2 \lambda u - 1}$	$\dfrac{\alpha - 1}{\alpha \cosh^2 \lambda u - 1}$		$\alpha - 1$	critical, limiting case of *ia*, when $k^2 \to 1$
ib'	$\beta < 1 = \alpha < x^{-1}$	$\dfrac{\sin h^2 \lambda u}{\cosh^2 \lambda u - \beta}$	$\dfrac{(1-\beta)\cos h\, \lambda u}{\cosh^2 \lambda u - \beta}$		$1 - \beta$	critical, limiting case of both *ib* and *ic*, when $k^2 \to 1$
ii	$\alpha \ne \beta$, imaginary $1 \le x^{-1}$	$\dfrac{\mathrm{sn}^2(\lambda u, k)}{1 - \mathrm{cn}^2 \lambda u + \lambda^2 \mathrm{cd}^2 \lambda u}$	$\dfrac{\lambda^2\left(1 - k^2\,\mathrm{sn}^2 \lambda u\,\mathrm{cd}^2 \lambda u\right)}{1 - \mathrm{cn}^2 \lambda u + \lambda^2 \mathrm{cd}^2 \lambda u}$	$\dfrac{1}{2}\left(1 - \dfrac{2 - \alpha - \beta}{2\sqrt{(1-\alpha)(1-\beta)}}\right)$	$\sqrt{(1-\alpha)(1-\beta)}$	revolution
iii	$\alpha = \beta = 1 < x^{-1}$	$\dfrac{u^2}{1 + u^2}$	$\dfrac{1}{1 + u^2}$			critical

Now putting $\beta_0 = j\beta_0'$, $\Gamma = j\Gamma'$ we have

$$P = -\beta_0' \Gamma'^{-1}$$

$$Q = \frac{1}{4} \Gamma'^{-1} \beta_1^2 \alpha_2^{-1} ,$$

$$\mu = j^2 \Gamma' \alpha_2^{-1} ;$$

from which we obtain

$$(\alpha, \beta) = \frac{1}{2} \beta_0' \Gamma'^{-1} \left(1 \pm \sqrt{1 - \Gamma' \beta_1^2 \beta_0'^{-2} \alpha_2^{-1}}\right).$$

From equation 8 we have

$$u = 2\alpha_2 \sqrt{\mu}\, \tau ,$$

where

$$\mu = j^2 \Gamma' \alpha_2^{-1} .$$

If $\Gamma' > \beta_0'^2 \alpha_2 \beta_1^{-2}$, α and β become imaginary so that this case corresponds to the case *ii*.

On the other hand, $\Gamma' \leq \beta_0'^2 \alpha_2 \beta_1^{-2}$ corresponds to case *i* or case *iii*. The subdivision within case *i* can be obtained easily

if

$\beta_0' - (\beta_1^2/4\alpha_2) < \Gamma' < \frac{1}{2}\beta_0'$, then case *ia*;

if

$0 < \Gamma' < \beta_0' - (\beta_1^2/4a_2)$, then case *ib*;

and if

$\Gamma' > \beta_0' - (\beta_1^2/4\alpha_2)$ and $\frac{1}{2}\beta_0'$, then case *ic*.

Rewriting the above and combining it with the real/imaginary criterion results in the following scheme:

Case A

$$\beta_0' > \frac{\beta_1^2}{2\alpha_2} \quad \text{or} \quad \frac{\beta_0'}{2} < \beta_0' - \frac{\beta_1^2}{4\alpha_2} < \frac{\beta_0'^2 \alpha_2}{\beta_1^2}$$

if

$0 < \Gamma' < \beta_0' - (\beta_1^2/4\alpha^2)$, then case *ib*;

if

$\beta_0' - (\beta_1^2/4\alpha_2) < \Gamma' < (\beta_0'^2 \alpha_2/\beta_1^2)$, then case *ic*;

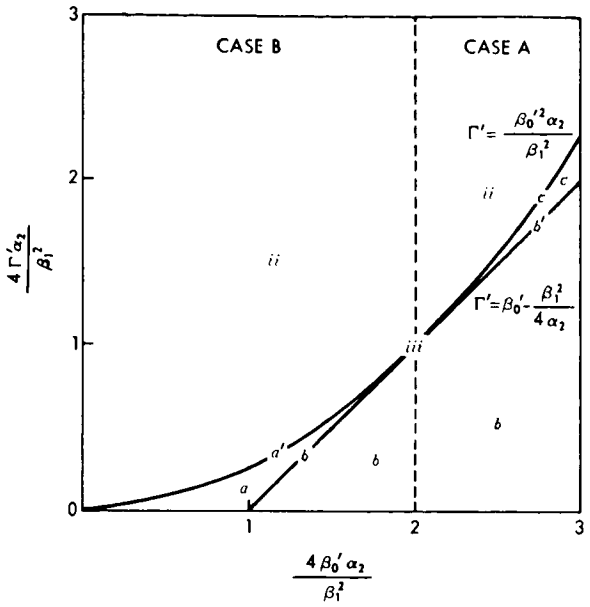


FIGURE 1.—Criterion for several cases, expressed by parameters Γ' and β_0' in units of (1/4) $(\beta_1^2 \alpha_2^{-1})$. The respective cases are described in table 1.

if

$(\beta_0'^2 \alpha_2/\beta_1^2) < \Gamma'$, then case *ii*.

Case B

$$\beta_0' < (\beta_1^2/2\alpha_2)$$

or

$$\beta_0' - (\beta_1^2/4\alpha_2) < (\beta_0'^2 \alpha_2/\beta_1^2) < (\beta_0'/2)$$

if

$0 < \Gamma' < \beta_0' - (\beta_1^2/4\alpha_2)$, then case *ib*;

if

$\beta_0' - (\beta_1^2/4\alpha_2) < \Gamma' < (\beta_0'^2 \alpha_2/\beta_1^2)$, then case *ia*;

if

$(\beta_0'^2 \alpha_2/\beta_1^2) < \Gamma'$, then case *ii*.

The critical cases are easily obtained by comparing the above scheme with table 1. In order to be able to distinguish between the different cases, a diagram is drawn with the parameters expressed by Γ' and β_0' in units of $\beta_1^2/4\alpha_2$ (fig. 1).

Now, the period of $\sin\omega_1$ in terms of τ is given by

$$T = \frac{4K}{2\alpha_2^{1/2}\, j\, \Gamma'^{1/2} \lambda} ,$$

TABLE 2.—*Behavior of the Period and the Character of the Motion*

Case	*ii*	*a'*	*a*	*b*	*b*	*b'*	*c*	*c*	*ii*
Period T	↗	∞	↘ $\frac{2\pi}{\sqrt{(\alpha-1)\Gamma'\cdot 2\alpha_2 j}}$ ↗			∞	↘ $\frac{2\pi}{\sqrt{(1-\alpha)\Gamma'\cdot 2\alpha_2 j}}$ ↗		
Character*	R	C	L	L	L	C	R	R	R
Remark		$\beta = \alpha$ $k = 1$		$\beta = 1$ $k = 0$		$\alpha = 1$ $k = 1$		$\beta = \alpha$ $k = 0$	

*The abbreviations are as follows: Revolution, Critical, and Libration.

where the complete elliptic integral of the first kind,

$$K = \int_0^{\pi/2} \frac{d\omega_1}{\sqrt{1 - k^2 \sin^2 \omega_1}} = \frac{1}{2}\pi F\left(\frac{1}{2}, \frac{1}{2}, 1, k^2\right). \tag{15}$$

Accordingly, if in figure 1 we take a point which is moving around the critical point *iii* in a counterclockwise direction, we will find the change of the period and of the character of the motion as given in table 2.

In any case the solutions of equations 1 up to the order considered here, are given in case *ia* by using equations 3, 7, 9, 12, and 13;

$$\cos^2 \omega = \sin^2 \omega_1 = \frac{\mathrm{sn}^2 \lambda u}{\alpha - \mathrm{cn}^2 \lambda u},$$

$$\rho = \mathrm{p} + \chi = \mathrm{p} + \mu^{1/2} \chi_0$$

$$= -\frac{1}{2} j \frac{\beta_1}{\alpha_2} \frac{\mathrm{sn}^2 \lambda u}{\alpha - \mathrm{cn}^2 \lambda u}$$

$$+ j \sqrt{\frac{\Gamma'}{\alpha_2}} \frac{(\alpha - 1)\, \mathrm{cn}\, \lambda u\, \mathrm{dn}\, \lambda u}{\alpha - \mathrm{cn}^2 \lambda u},$$

where

$$\alpha = \frac{1}{2} \beta_0' \Gamma'^{-1} \left(1 + \sqrt{1 - \Gamma' \beta_1^2 \beta_0'^{-2} \alpha_2^{-1}}\right),$$

$$k^2 = \frac{\beta - 1}{\alpha - 1}, \qquad \lambda u = 2\sqrt{\alpha - 1}\, \alpha_2^{1/2} \Gamma'^{1/2} j\tau.$$

For the other cases, we can easily obtain the results from table 1 in a similar manner.

THE CASE OF MODERATE ECCENTRICITY

In this case, we expand α and β into power series in j. From the expressions of equation 9 it

follows, by neglecting high order terms, that

$$\left.\begin{aligned} \alpha &= \frac{\beta_0}{\Gamma} - j\left(\frac{1}{4}\frac{\beta_1^2}{\alpha_2\beta_0} - \frac{\gamma_0}{\beta_0}\right), \\ \beta &= j\left(\frac{1}{4}\frac{\beta_1^2}{\alpha_2\beta_0} - \frac{\gamma_0}{\beta_0} - \frac{\beta_2}{\alpha_2} + \frac{3}{2}\frac{\alpha_3\beta_1}{\alpha_2^2}\right). \end{aligned}\right\} \tag{16}$$

Here, except for the critical case, we have only one of two cases: *ib* or *ic*

Case ib: *Libration*

From equations 8 and 10 we obtain

$$\frac{du}{1 + \frac{\alpha_3}{\alpha_2}\mu^{1/2}\chi_0 + \left(\frac{3}{2}\frac{\alpha_4}{\alpha_2} - \frac{7}{8}\frac{\alpha_3^2}{\alpha_2^2}\right)\mu\chi_0^2} = 2\alpha_2\mu^{1/2}\,d\tau, \tag{17}$$

from which, expanding it into a power series in $\mu^{1/2}$, we obtain

$$du\left[1 - \frac{\alpha_3}{\alpha_2}\mu^{1/2}\chi_0 + \left(-\frac{3}{2}\frac{\alpha_4}{\alpha_2} + \frac{15}{8}\frac{\alpha_3^2}{\alpha_2^2}\right)\mu\chi_0^2\right] = 2\alpha_2\mu^{1/2}\,d\tau, \tag{18}$$

up to the order $\mu = 0(j)$, which corresponds to j^2 in equation 1. We also have, from equations 7 and 8,

$$\chi = \frac{\alpha_2}{a_2}\mu^{1/2}\chi_0\left[1 - \frac{1}{2}\frac{\alpha_3}{\alpha_2}\mu^{1/2}\chi_0 + \left(-\frac{1}{2}\frac{\alpha_4}{\alpha_2} + \frac{5}{8}\frac{\alpha_3^2}{\alpha_2^2}\right)\mu\chi_0^2\right].$$

Now, putting

$$\left.\begin{aligned} \sigma_1 &= -\frac{\alpha_3}{\alpha_2}, \\ \sigma_2 &= -\frac{3}{2}\frac{\alpha_4}{\alpha_2} + \frac{15}{8}\frac{\alpha_3^2}{\alpha_2^2}, \end{aligned}\right\} \tag{19}$$

into equation 18 and using $\chi_0 = cn\lambda u$ to the required order, we have

$$\begin{aligned} du\left(1 + \sigma_1\mu^{1/2}\,\mathrm{cn}\,\lambda u + \sigma_2\mu\,\mathrm{cn}^2\lambda u\right) &= 2\alpha_2\mu^{1/2}\,d\tau \\ &\equiv du_1. \end{aligned}$$

It is known that (reference 9, pp. 58 and 62)

$$\int_0^v \mathrm{cn}\,v\,dv = k^{-1}\sin(k\,\mathrm{sn}\,v), \qquad 0 \le k^2 \le 1,$$

$$\int_0^v \mathrm{cn}\,v^2\,dv = k^{-2}\left[E(v) - k'^2 v\right],$$

where $E(v)$ is the elliptic integral of the second kind such as

$$E(v) = \int_0^v dn^2\, v\, dv \ .$$

Accordingly,

$$\lambda u_1 = \lambda\left(1+\sigma_2\mu s\right)u + \sigma_1\mu^{1/2}k^{-1}\sin^{-1}(k\ sn\,\lambda u) + k^{-2}\sigma_2\mu Z(\lambda u) \ , \tag{20}$$

$$\begin{aligned} Z(v) &= E(v) - EK^{-1}v \ , \\ s &= 1 + k^{-2}\left(EK^{-1} - 1\right) \ , \\ E &= \int_0^{\pi/2} \left(1 - k^2\sin^2 v\right)^{1/2} dv \\ &= \frac{\pi}{2} F\left(-\frac{1}{2}, \frac{1}{2}; 1; k^2\right) \ , \end{aligned} \tag{21}$$

and K is given by equation 15.

Equation 20 has a form

$$\lambda u_1 = \lambda(1+\sigma_2\mu s)u + (\text{periodic terms in } \lambda u \text{ with the period } 4K).$$

Therefore, by inversion, it follows that

$$\lambda u = \lambda_1 u_1 + (\text{periodic terms in } \lambda_1 u_1),$$

where

$$\lambda_1 = \lambda(1+\sigma_2\mu s)^{-1}.$$

Thus

$$\begin{aligned} \lambda u &= \nu\tau - \sigma_1\sqrt{\mu}\,k^{-1}\sin^{-1}\left(k\ sn\,\nu\tau\right) \\ &\quad + \sigma_1^2\mu\,cn\,\nu\tau\,k^{-1}\sin^{-1}(k\ sn\,\nu\tau) - \sigma_2\mu k^{-2}Z(\nu\tau) \ , \end{aligned} \tag{22}$$

where

$$\begin{aligned} \nu &= 2a_1\sqrt{\mu}\,\lambda_1 \\ &= 2a_1^{1/2}\,j^{1/2}\,\Gamma^{1/2}\,(\alpha-\beta)^{1/2}\left(1-\sigma_2\mu s\right) \ , \end{aligned}$$

and $k^2 = (1-\beta)(\alpha-\beta)^{-1}$. The other notations, σ_1 and σ_2, are given by equations 19; s by equation 21;

and α, β by equations 16. Then,

$$x = \cos^2\omega = \frac{\mathrm{sn}^2(\lambda u, k)}{\alpha - \beta\,\mathrm{cn}^2(\lambda u, k)}$$

$$X = \sqrt{\frac{j\Gamma}{\alpha_2}}\left\{\left[1 - \frac{j\Gamma}{\beta_0}\left(\frac{1}{4}\frac{\beta_1^2}{\alpha_2\beta_0} - \frac{\gamma_0}{\beta_0}\right)\right]\mathrm{cn}\,\lambda u - \frac{1}{2}\frac{\alpha_3}{\alpha_2}\sqrt{\frac{j\Gamma}{\alpha_2}}\,\mathrm{cn}^2\lambda u\right.$$
$$\left. + \frac{j\Gamma}{\beta_0}\left(\frac{1}{4}\frac{\beta_1^2}{\alpha_2\beta_0} - \frac{\gamma_0}{\beta_0} - \frac{1}{2}\frac{\alpha_4\beta_0}{\alpha_2^2} + \frac{5}{8}\frac{\alpha_3^2\beta_0}{\alpha_2^2}\right)\mathrm{cn}^3\lambda u\right\}.$$

Therefore, we have, neglecting higher orders, the following forms as the solutions of equation 1,

$$\cos\omega = \left(\frac{\Gamma}{\beta_0}\right)^{1/2}\left[1 + \frac{j\Gamma}{2\beta_0}\left(\frac{1}{2}\frac{\beta_1^2}{\alpha_2\beta_0} - \frac{2\gamma_0}{\beta_0} - \frac{\beta_2}{\alpha_2} + \frac{3}{2}\frac{\alpha_3\beta_1}{\alpha_2^2}\right)\right]\mathrm{sn}\,\lambda u$$
$$- \left(\frac{\Gamma}{\beta_0}\right)^{1/2}\frac{j\Gamma}{2\beta_0}\left(\frac{1}{4}\frac{\beta_1^2}{\alpha_2\beta_0} - \frac{\gamma_0}{\beta_0} - \frac{\beta_2}{\alpha_2} + \frac{3}{2}\frac{\alpha_3\beta_1}{\alpha_2^2}\right)\mathrm{sn}^3\lambda u\,,$$

$$\rho = \sqrt{\frac{j\Gamma}{\alpha_2}}\left[1 - \frac{j\Gamma}{\beta_0}\left(\frac{1}{4}\frac{\beta_1^2}{\alpha_2\beta_0} - \frac{\gamma_0}{\beta_0}\right)\right]\mathrm{cn}\,\lambda u - \frac{1}{2}\frac{\alpha_3\Gamma}{\alpha_2^2}\,j\,\mathrm{cn}^2\lambda u$$
$$- \frac{1}{2}\frac{\beta_1\Gamma}{\alpha_2\beta_0}\,j\,\mathrm{sn}^2\lambda u + \sqrt{\frac{j\Gamma}{\alpha_2}}\frac{j\Gamma}{\beta_0}\left(\frac{1}{4}\frac{\beta_1^2}{\alpha_2\beta_0} - \frac{\gamma_0}{\beta_0}\right.$$
$$\left. - \frac{1}{2}\frac{\alpha_4\beta_0}{\alpha_2^2} + \frac{5}{8}\frac{\alpha_3^2\beta_0}{\alpha_2^2\alpha_2}\right)\mathrm{cn}^3\lambda u\,,$$

where λu is given by equation 22. The order of magnitude adopted here corresponds to the second order of equation 7.

Case ic: *Revolution*

In this case a formula similar to the aforementioned one can be given:

$$du\left(1 + \sigma_1\mu^{1/2}\,\mathrm{dn}\,\lambda u + \sigma_2\mu\,\mathrm{dn}^2\lambda u\right) = d\bar{u}_1\,,$$

from which we have

$$\lambda u_1 = \lambda u + \sigma_1\mu^{1/2}\,\mathrm{am}\,\lambda u + \sigma_2\mu\,\mathbf{E}(\lambda u)$$
$$= \left(1 + \sigma_1\mu^{1/2}\frac{\pi}{2}\mathbf{K}^{-1} + \sigma_2\mu\,\mathbf{E}\mathbf{K}^{-1}\right)\lambda u \qquad (23)$$
$$+ \sigma_1\mu^{1/2}\left(\mathrm{am}\,\lambda u - \frac{\pi}{2\mathbf{K}}\lambda u\right) + \sigma_2\mu\,\mathbf{k}^{-2}\,\mathbf{Z}(\lambda u)\,.$$

Putting

$$\nu = 2\alpha_2 \mu^{1/2} (1-\beta)^{1/2} \left(1 + \sigma_1 \mu^{1/2} \frac{\pi}{2} K^{-1} + \sigma_2 \mu E K^{-1}\right)^{-1} ,$$

we have, by inversion,

$$\lambda u = \nu\tau - \sigma_1 \mu^{1/2} \left(\mathrm{am}\,\nu\tau - \frac{\pi}{2} K^{-1} \nu\tau\right)$$

$$+ \sigma_1^2 \mu \left(\mathrm{dn}\,\nu\tau - \frac{\pi}{2} K^{-1}\right)\left(\mathrm{am}\,\nu\tau - \frac{\pi}{2} K^{-1} \nu\tau\right) - \sigma_2 \mu Z(\nu\tau) ,$$

where

$$k^2 = (\alpha - \beta)(1-\beta)^{-1} .$$

We note that $\sigma_1\mu^{1/2}am\lambda u$, the second term of the right side of equation 23, also contributes to the change of the period-contrary to case *ib*, where the corresponding term, $\sigma_1\mu^{1/2}k^{-1}\sin^{-1}(ksn\lambda u)$, is purely periodic provided $0 \leq k^2 \leq 1$. It is easily seen that, roughly speaking, case *c* corresponds to case *b* where $k > 1$. At any rate, for the former case we have the following solutions of equation 1:

$$\cos\omega = \sin\omega_1 = \mathrm{sn}\,\lambda u \left[1 + \frac{1}{2} j \left(\frac{1}{4} \frac{\beta_1^2}{\alpha_2 \beta_0} - \frac{\gamma_0}{\beta_0} - \frac{\beta_2}{\alpha_2} + \frac{3}{2} \frac{\alpha_3 \beta_1}{\alpha_2^2}\right) \mathrm{cn}^2\,\lambda u\right] ;$$

$$\rho = \mathrm{p} + \chi = \sqrt{\frac{j\Gamma}{\alpha_2}}\,\mathrm{dn}\,\lambda u - \frac{1}{2} j \frac{\beta_1}{\alpha_2} \mathrm{sn}^2\,\lambda u - \frac{1}{2} j \frac{\Gamma\alpha_3}{\alpha_2^2} \mathrm{dn}^2\,\lambda u$$

$$- j\sqrt{\frac{j\Gamma}{\alpha_2}} \left(\frac{1}{4} \frac{\beta_1^2}{\alpha_2 \beta_0} - \frac{\gamma_0}{\beta_0}\right) \mathrm{sn}^2\,\lambda u\,\mathrm{dn}\,\lambda u + \left(\frac{j\Gamma}{\alpha_2}\right)^{3/2} \left(-\frac{1}{2} \frac{\alpha_4}{\alpha_2} + \frac{5}{8} \frac{\alpha_3^2}{\alpha_2^2}\right) \mathrm{dn}^3\,\lambda u .$$

GENERAL SCHEME OF THE FURTHER APPROXIMATIONS

In order to solve the equations with higher order terms, it is first necessary to solve an equation such as equation 6 which contains these terms. Even if the original equations of motion are limited to finite terms, the equation 6-type of equation is, in general, an algebraic equation with $\cos^2\omega$ as the unknown. When combined with the second equation of equation 4, we find that $(d/d\tau)(\cos^2\omega)$ is equal to an algebraic function in $\cos^2\omega$. The solution of a differential equation of this kind would involve Abelian integrals—many valued functions, more complicated than the elliptic integrals. Therefore, the solution becomes so complicated that we cannot see the general character of the motion at a first glance. However, if we restrict ourselves to the solution of a real function, the complicacy would be much reduced.

In this section we are concerned with the cases where the character of motion is quite similar to the former cases, except for slight changes caused by the higher order terms.

Now, we consider the equations of motion in the form

$$d\chi/d\tau = -\partial\Psi/\partial\omega_1,$$

$$d\omega_1/d\tau = \partial\Psi/\partial\chi,$$

where

$$\Psi = j a_0 \left(\sin^2\omega_1, j\right) + \sum_2 a_n \left(\sin^2\omega_1, j\right) \chi^n$$

and $a_n (n = 2, 3, \cdots)$ is a power series in j,

$a_2(\sin^2\omega_1, 0) = a_{2,0}$ and is a constant which corresponds to the previous case if $a_{2,0} = \alpha_2$.

From the integral $\Psi - j\Gamma = 0$, we obtain

$$4a_2^2\chi^2\left(1 + \Sigma a_n^* \chi^{n-2}\right) = 4a_2 j\left(\Gamma - a_0\right) \equiv \eta^2 ,$$

from which

$$2a_2\chi = \eta + \sum_1 b_n \eta^n .$$

Thus, we have

$$(\partial\Psi/\partial\chi)^{-1}\, d\omega_1 = d\tau$$

or

$$\eta^{-1}\left(1 + \Sigma c_n \eta^n\right)^{-1} d\omega_1 = \text{constant} \cdot d\tau .$$

If the solution of $\eta^{-1}\, d\omega_1 = du^*$ could be obtained in a simple way, as in the previous cases, then using u^* as an intermediary parameter would be a direct extension of the previous cases. But this integral is, in general, an Abelian integral as we have just stated. A general treatment is not suitable for practical purposes. As an alternative, we shall consider the following situations.

α: Case of moderate eccentricity

Let $a_0(\sin^2\omega_1, 0) = a_{0,1} \sin^2\omega_1$, $a_{0,1} > 0$; then the solution of $a_0(\sin^2\omega_1, 0) - \Gamma = 0$ is

$$\sin^2\omega_1 = \left(a_{0,1}\right)^{-1}\Gamma . \tag{24}$$

Now, we denote one of the solutions by α^{-1}:

$$a_0\left(\sin\omega_1, j\right) - \Gamma = 0 ,$$

the first approximation of which is equation 24. We then have

$$4a_2 j\left[\Gamma - a_0\left(\sin^2\omega_1, j\right)\right] = \mu_1\left(1 - \alpha\sin^2\omega_1\right)\left(1 + \Sigma A_n \sin^{2n}\omega_1\right) ,$$

where $\mu_1 = 4\Gamma a_{2,0} j$. Therefore,

$$\eta = \sqrt{\mu_1}\,\xi\left(1 + \Sigma\beta_n \sin^{2n}\omega_1\right) ,$$

where $\xi^2 = 1 - \alpha\sin^2\omega_1$ and

$$\xi^{-1}\left[1 + \Sigma\mu_1^{n/2} f_n\left(\sin^2\omega_1\right)\xi^n\right]^{-1} d\omega_1 = \mu_1^{1/2}\, d\tau .$$

Now

$$\xi^{-1}\, d\omega_1 = du . \tag{25}$$

Here we again divide the problem into two subcases:

Case ib: $\alpha > 1$

We have

$$\sqrt{\alpha}\sin\omega_1 = \operatorname{sn}\left(\lambda u, \alpha^{-1/2}\right) ,$$

$$\xi = \sqrt{1 - \alpha\sin^2\omega_1}$$

$$= \operatorname{cn}\left(\lambda u, \alpha^{-1/2}\right)$$

$$\lambda = \alpha^{1/2} ;$$

then

$$\left[1 + \Sigma\mu^{n/2} f_n\left(\alpha^{-1}\operatorname{sn}^2\alpha^{1/2}u\right)\left(\operatorname{cn}^n\alpha^{1/2}u\right)\right]^{-1} du = \mu_1^{1/2}\, d\tau$$

or

$$\left[1 + \Sigma \mu^{n/2} g_n \left(\alpha^{-1} \operatorname{sn}^2 \alpha^{1/2} u\right) \operatorname{cn}^n \alpha^{1/2} u\right] du = \mu_1^{1/2} d\tau , \tag{26}$$

where g_n is a power series in j with the coefficients polynomials in $\alpha^{-1}sn^2\lambda u$. Therefore, the integration of equation 26 is expressed as

$$u + \Sigma B_n \int \operatorname{cn}^n \lambda u \, du = \mu_1^{1/2} \tau , \tag{27}$$

where B_n is a numerical constant. This integration is carried out by the use of the Jacobian elliptic function and the Jacobian zeta-function (see reference 10, p. 164). Thus, the final form is expressed by

$$\lambda u + \text{a periodic function of } \lambda u = \text{constant} \cdot \tau,$$

where the constant factor in the right side is $\mu_1^{1/2}\lambda[1+0(j)]$. This is an extension of equation 20.

Case ic: $\alpha < 1$

From equation 25 we have

$$\sin \omega_1 = \operatorname{sn}\left(u, \alpha^{1/2}\right) ,$$

$$\xi = \sqrt{1 - \alpha \sin^2 \omega_1} = \operatorname{dn}\left(u, \alpha^{1/2}\right);$$

therefore,

$$\left[1 + \Sigma \mu^{n/2} f_n \left(\operatorname{sn}^2 u\right) \operatorname{dn}^n u\right]^{-1} du = \mu_1^{1/2} d\tau$$

or

$$\left[1 + \Sigma \mu^{n/2} g_n \left(\operatorname{sn}^2 u\right) \operatorname{dn}^n u\right] du = \mu_1^{1/2} d\tau .$$

In this case we have, instead of equation 27,

$$u + \Sigma C_n \int \operatorname{dn}^n u \, du = \mu_1^{1/2} \tau ,$$

where the integration will be carried out, as above, by using the Jacobian elliptic function and the Jacobian zeta-function. Thus we have

$$u + a \text{ periodic function of } u = \text{constant} \cdot \tau,$$

where the constant factor in the right side again is $\mu_1^{1/2}[1+0(j)]$.

β: Case of small eccentricity

In this case the most important factor in $\Gamma - a_0$ $(\sin^2\omega_1, j)$ is not the $1-\alpha \sin^2\omega_1$ term, as in the previous case, but is $(1-\alpha \sin^2\omega_1)(1-\beta \sin^2\omega)$, where both α and β are of the order of 1. Therefore we cannot expand each of the above factors into power series beginning with a constant term.

Or more concretely, hereafter $a_{0,1}$ is assumed to be of the order of j: $a_{01} = ja'_{0,1}$ and, also, it is assumed that

$$\lim_{j \to 0} j^{-1} a_0 \left(\sin^2 \omega_1, j\right) = a'_{0,1} \sin^2 \omega_1 + a_{0,2} \sin^4 \omega_1 ,$$

where $a'_{0,1}$ corresponds to β_0^1, and $a_{0,2}$ to $-(1/4)\beta_1^2\alpha_2^{-1}$. Also, we denote by α^{-1}, β^{-1} the solutions of

$$a_0 \left(\sin^2 \omega_1, j\right) - j\Gamma' = 0 ,$$

which approach the solutions of

$$\Gamma' - a'_{0,1} \sin^2 \omega_1 - a_{0,2} \sin^4 \omega_1 = 0 .$$

As before, we have

$$4a_2 j \left[j\Gamma' - a_0 \left(\sin^2 \omega_1 , j\right)\right] = \mu_1 \left(1 - \alpha \sin \omega_1\right) \left(1 - \beta \sin^2 \omega_1\right) \left(1 + \Sigma A_n \sin^{2n} \omega_1\right) ,$$

where $\mu_1 = 4\Gamma' a_{2,0} j^2$. Therefore,

$$\eta = \mu_1^{1/2} \chi_0 \left(1 + \Sigma \beta_n \sin^{2n} \omega_1\right) ,$$

$$\chi_0^2 = \left(1 - \alpha \sin^2 \omega_1\right) \left(1 - \beta \sin^2 \omega_1\right)$$

and, as before,

$$\chi_0^{-1}\left[1+\Sigma\mu_1^{n/2}\, f_n\left(\sin^2\omega_1\right)\chi_0^n\right]^{-1} d\omega_1 = \mu_1^{1/2}\, d\tau\ . \tag{28}$$

Since $\chi_0^{-1}\, d\omega_1 = du$, we have, for case i (except for the critical cases):

$$x = \sin^2\omega_1 = \frac{sn^2(\lambda u, k)}{p - q\, cn^2(\lambda u_1 k)}, \tag{29}$$

where λ, k, p, and q can be obtained in table 1, according to the respective cases. Further,

$$\chi_0 = \frac{(p-q)\, cn^{\epsilon_1}\lambda u\, dn^{\epsilon_2}\lambda u}{p - q\, cn^2(\lambda u_1 k)}, \tag{30}$$

where ϵ_1, $\epsilon_2 = 0$ or 1, but both are not zero at the same time.

Now, from equation 28 we have

$$\left[1 + \Sigma\mu_1^{n/2}\, f_n(x)\,\chi_0^n\right]^{-1} du = \mu_1^{1/2}\, d\tau$$

or

$$\left[1+\Sigma\mu_1^{n/2}\, g_n(x)\,\chi_0^n\right] du = \mu_1^{1/2}\, d\tau\ , \tag{31}$$

where g_n is a polynomial of x. Thus, for n even, g_n has the form

$$g_n = G_1\left(sn^2\lambda u\right)\ , \tag{32}$$

where G_1 denotes a rational function of the argument. The integration of equation 32 produces an elliptic integral of the third kind, in general, unlike Case α (reference 10, p. 164). On the other hand, for n odd,

$$g_n = G_2\left(sn^2\lambda u\right)\, cn^{\epsilon_1}\lambda u\, dn^{\epsilon_1}\lambda u\ ,$$

the integration of which requires only the Jacobian elliptic functions.

For case *ii*, however, the equations corresponding to equations 29 and 30 have the different forms; but all the functions g_n have a single form such as equation 32.

In any case, we have, after integration of equations 28 or 32, the form

λu + a periodic function of λu = constant $\cdot\, \tau$,

from which we obtain

λu = constant $\cdot\, \tau$

+ a periodic function of constant $\cdot\, \tau$.

It should be noted that the odd power terms in equation 31 do not contribute any change to the period of libration, where $\epsilon_1 = 0$, which corresponds to the case of libration.

DISCUSSION

The denominator of the right side of equation 11 is a square root of a quartic of x. A general treatment of such a case was carried out by Andoyer[11] and extended by Hagihara[12] in connection with the libration problem of asteroids (see also reference 13). Both of them used the ζ-function of Weierstrass; nevertheless, we feel the integration form described here is more suitable for this special case, owing to the fact that only real functions are involved. For example, if $\beta \to 0$ in equation 10,

$$\frac{d\omega_1}{\sqrt{1 - k^2\sin^2\omega_1}} = du\ ,$$

where $k^2 = \alpha$, from which, if $0 < \alpha < 1$, we have $\sin\omega_1 = snu$, which is, of course, a limiting case of *ic*, etc. Thus the present work represents some natural extensions of the derivation of well-known Jacobian elliptic functions.

For the numerical computation, integrating equation 17 in the literal form may not necessarily be the best way, for a direct numerical integration of equation 17 is more effective, since its denominator generally does not become very small—otherwise, the numerical integration would lose its validity. However, the problem of inversion of a numerical function still would remain.

CONCLUSION

Equations 1, which have terms up to second order of j, have been solved both for the case of small eccentricity and for moderate eccentricity. In the former, the intermediary solutions themselves are sufficient for the present discussion, since the solutions have an accuracy of the order of 1 for $\cos\omega$ (neglecting the order of j) and of the order of j for ρ (neglecting j^2). This accuracy corresponds to the order of j^2 in the Hamiltonian, since ρ has the factor of j and the argument of the solution is neither τ alone nor multiplied by a constant factor of order of 1, but τ multiplied by that of order j.

On the other hand, in the second or normal cases, the solutions have the accuracy of order j for $\cos\omega$ and $j^{3/2}$ for ρ, which corresponds to j^2 in the Hamiltonian.

ACKNOWLEDGMENTS

The author would like to express his gratitude to Mr. I. G. Izsak and to Drs. P. Musen and Y. Kozai for their discussions.

REFERENCES

1. Hori, G. I., "The Motion of an Artificial Satellite in the Vicinity of the Critical Inclination," *Astronom. J.*, **65**(5): 291–300, June 1960.
2. Garfinkel, B., "On the Motion of a Satellite in the Vicinity of the Critical Inclination," *Astronom. J.*, **65**(10): 624–627, December 1960.
3. Hagihara, Y., "Libration of an Earth Satellite With Critical Inclination," *Smithsonian Contrib. to Astrophys.*, **5**(5): 39–51, 1961.
4. Kozai, Y., "Motion of a Particle With Critical Inclination in the Gravitational Field of a Spheroid," *Smithsonian Contrib. to Astrophysics*, **5**(5): 53–58, 1961.
5. Izsak, I. G., "On the Critical Inclination in Satellite Theory," Smithsonian Inst., Astrophys. Observ., Res. in Space Sci., Spec. Rept. No. 90, Mar. 14, 1962.
6. Neville, E. H., "Jacobian Elliptic Functions," 2d ed., Oxford: Clarendon Press, 1951.
7. Whittaker, E. T., and Watson, G. N., "A Course of Modern Analysis," 4th ed., Cambridge: University Press, 1927.
8. Kozai, Y., "Second-Order Solution of Artificial Satellite Theory without Air Drag," *Astronom. J.*, **67**(7): 446–461, September 1962.
9. Hancock, H., "Elliptic Integrals," New York: Dover, 1958.
10. Grobner, W., and Hofreiter, N., "Integraltafel. vol. 1, Unbestimmte Integrale," 2d ed., Vienna: Springer-Verlag, 1957.
11. Andoyer, H., "Contribution a la Theorie des Petites Planetes dont le Moyen Mouvement est Sensiblement Double de celui de Jupiter," *Bulletin Astronomique*, **20**, 321–356, September 1903.
12. Hagihara, Y., "On the General Theory of Libration," *Japan. J. Astron. and Geophys.*, **21**(1–2): 29–43, 1944.
13. Halphen, G. H., "Traite des Fonctions Elliptiques et de leurs Applications," Paris: Gauthier-Villars, 1886, vol. 1, chs. 4 and 5.

Appendix A

LIST OF SYMBOLS

c_d, c_n, d_n	Jacobian elliptic functions (see pp. 154 and 157)
e_f, e_g, e_h	Parameter of Weierstrassian elliptic function (see p. 153)
E	Complete elliptic integral of the second order (see p. 160)
f_j, g_j, h_j	First order elliptic function associated with the $\wp$-function (see reference 6 and p. 153)
J_n	Harmonic coefficient of the earth's potential (see p. 150)
k	Modulus of the Jacobian elliptic function (see pp. 153 and 156)
K	Complete elliptic integral of the first order (see p. 158)
$\wp$	Weierstrassian elliptic function (see p. 153)
s_n	Jacobian elliptic function (see p. 154)
t	Time
T	Period of the solutions (see p. 157)
u	u will equal t, except for the constant factor, if higher orders are neglected (see p. 157)

$x = \sin^2\omega_1 = \cos^2\omega$

α, β, α_i, β_i, γ_i	Constants
$\Gamma = j\Gamma'$	Energy constant (see p. 151)
λ	Constant factor in the argument of the Jacobian elliptic functions (see p. 157)
ρ	Variable related to the eccentricity (see reference 5 and p. 150)
ω	Mean argument of the perigee

$\omega_1 = (\pi/2) - \omega$

ω_f, ω_g, ω_h	Semi-periods associated with the Weierstrassian elliptic function (see p. 153)

AN ANALYTICAL REPRESENTATION OF MUSEN'S THEORY OF ARTIFICIAL SATELLITES IN TERMS OF THE ORBITAL TRUE LONGITUDE

ANN E. BAILIE AND DAVID FISHER

Goddard Space Flight Center

Analytic solutions to order k_2 of the differential equations of Musen's Theory of Artificial Satellites are presented. These solutions include long-period terms and terms with small divisors derived from the k_2^2 approximation as well as from the third and fourth harmonics of the earth's potential. The results obtained herein may be used for computing orbits of artificial satellites to the first order, for checking the machine programming of the numerical theory, and for comparing Musen's theory with other artificial satellite theories. In addition, these results may be incorporated into the numerical theory to broaden its applicability, particularly in the case of nearly circular orbits.

INTRODUCTION

Musen's theory of artificial satellites[1] in terms of the orbital true longitude permits numerical calculation of the perturbations of a satellite's orbit due to the earth's potential field. The degree of accuracy is limited only by the knowledge of the earth's potential field and the capabilities of the computer. This theory is also a useful tool for improving our knowledge of the geodetic parameters of the earth's potential. However, any numerical theory has the disadvantage that the characteristics of the solution—such as its behaviour for nearly circular orbits or in the vicinity of the critical angle of 63.4°—are not explicitly given. In addition, comparison with other satellite theories is virtually impossible.

For these reasons the development of an analytical solution of Musen's theory was undertaken. Since the theory has been adopted by NASA for orbit computation in its numerical form, an analytical solution is particularly useful—first as a check for the machine program, and secondly as a first approximation to the numerical series. For example, substituting the analytical expression of $(1-c)$, the motion of the argument of perigee in the plane of the orbit, to order k_2^2, into the numerical program would eliminate a division by e_0 which would cause trouble in the case of a nearly circular orbit.

The basic arguments in this report are expressed in terms of the orbital true longitude v. They are the true anomaly of the auxiliary ellipse,

$$\xi = cv - \pi_0 = v - [(1-c)v + \pi_0];$$

and the argument of latitude of the auxiliary satellite,

$$\eta = gv - \theta_0 = v - [(1-g)v + \sigma_0] .$$

Since the departure point is arbitrary, σ_0 is taken equal to θ_0. All other nomenclature has been kept the same as in Musen's paper[1]:

- a the semimajor axis,
- e the eccentricity,
- $h = [a(1-e^2)]^{-1/2}$
- i the inclination of the orbit plane to the Earth's equator,
- $n = a^{-3/2}$, the mean motion,
- $p = a(1-e^2)$.
- θ the right ascension of the ascending node
- σ the angular distance of the node from the departure point X,
- φ the elliptic $cv - \pi_0$.

PERTURBATIONS IN THE ORBIT PLANE

The derivative of Hansen's W function in terms of the orbital true longitude as given by Musen is

$$\frac{dW}{dv} = h_0^2 \left(\frac{1}{u^2}\frac{\partial\Omega}{\partial v}\right)\left\{2\frac{u}{h_0^2}\cos(\xi-\varphi) - 1 - e_0\cos\varphi + 2\left(\frac{h}{h_0}\right)^2\left[\cos(\xi-\varphi) - 1\right]\right\}\frac{h}{h_0} - 2\frac{\partial\Omega}{\partial u}\frac{h}{h_0}\sin(\xi-\varphi) + (1-c)\left[\frac{\partial W}{\partial\varphi} - \left(1+\frac{h_0}{h}\right)e_0\sin\varphi\right].$$

For the first approximation of W, we let

$$u = \bar{u} = h_0^2(1 + e_0\cos\xi),$$

$$\frac{h}{h_0} = \frac{h_0}{h} = 1,$$

$$\frac{\partial W}{\partial\varphi} = 0,$$

$$\lambda_1 = \sin\frac{i_0}{2},$$

$$\lambda_4 = \cos\frac{i_0}{2},$$

$$\lambda_2 = \lambda_3 = 0.$$

The disturbing function is taken to be the same as that given by Musen:

$$\Omega = k_2u^3(1-3\psi^2) + k_3u^4(3\psi-5\psi^3) + k_4u^5(3-30\psi^2+35\psi^4).$$

After substituting this into equation 1 and expanding, we develop a trigonometric series for dW/dv. The quantity $(1-c)$ is determined in such a way that no $\sin\varphi$ term appears in dW/dv. The first approximation to $(1-c)$ is

$$(1-c) = -\frac{3}{2}k_2h_0^4\left(1 - 3\cos^2 i_0\right).$$

From the second approximation of dW/dv we determine the value of this secular term to order k_2^2:

$$
\begin{aligned}
c = 1 &+ \frac{3}{2} k_2 h_0^4 \left(1 - 3\cos^2 i_0\right) \\
&- \frac{15}{16} k_4 h_0^8 \left(4 + 3e_0^2\right)\left(3 - 30\cos^2 i_0 + 35\cos^4 i_0\right) \\
&+ \frac{3}{32} k_2^2 h_0^8 \left[\left(162 - 64\sqrt{1 - e_0^2} + 25e_0^2\right)\right. \\
&\quad - \left(1060 - 480\sqrt{1 - e_0^2} + 74e_0^2\right)\cos^2 i_0 \\
&\quad \left. + \left(1650 - 864\sqrt{1 - e_0^2} + 9e_0^2\right)\cos^4 i_0\right] .
\end{aligned} \tag{2}
$$

The integration of dW/dv yields

$$
W = C_0 + C_1 \cos\varphi + \sum A' \cos\left(i\xi + j\varphi + 2k\eta\right) + \sum B' \sin\left[i\xi + j\varphi + (2k + 1)\eta\right] . \tag{3}
$$

The constants of integration C_0 and C_1 are determined from $dn_0\,\delta z/dv$ in such a way that no constant and no $\cos\xi$ term should appear in that derivative. Finally $\bar{W}$ (the function needed to obtain $n_0\,\delta z$, the perturbations in the mean anomaly; and $(1+\nu)$, the perturbations in the radius vector) is itself obtained from W by setting $\varphi = \xi$:

$$
\bar{W} = C_0 + C_1 \cos\xi + \sum A \cos\left[i\xi + 2k\eta\right] + \sum B \sin\left[i\xi + (2k + 1)\eta\right] . \tag{4}
$$

The perturbations in the mean anomaly are developed from the derivative[1]

$$
\frac{dn_0\,\delta z}{dv} = \left(1 - e_0^2\right)^{\frac{3}{2}} \left(\frac{h_0^2}{\bar{u}}\right)^2 \frac{\frac{h_0^2}{\bar{u}}\bar{W} + \nu^2}{1 + \frac{h_0^2}{\bar{u}}\bar{W}} ; \tag{5}
$$

or, in a more convenient form,

$$
\frac{dn_0\,\delta z}{dv} = \left(1 - e_0^2\right)^{\frac{3}{2}} \left(\frac{\bar{r}}{p_0}\right)^3 \bar{W} + \left(1 - e_0^2\right)^{\frac{3}{2}} \left(\frac{\bar{r}}{p_0}\right)^2 \nu^2 - \left(\frac{\bar{r}}{p_0}\right) \bar{W} \frac{dn_0\,\delta z}{dv} . \tag{6}
$$

where

$$
h_0^2/\bar{u} = \bar{r}/p_0 .
$$

For the first approximation of $dn_0\ \delta z/dv$ we let

$$\nu=0, \qquad dn_0\ \delta z/dv=0$$

in the right-hand side of equation 6, and we have

$$\left[\frac{dn_0\delta z}{dv}\right]_1 = \left(1-e_0^2\right)^{\frac{3}{2}}\left(\frac{\bar{r}}{p_0}\right)^3 \bar{W}. \tag{7}$$

In the second approximation, equation 7 is used in the last term of equation 6 and ν^2 is obtained from $\bar{u}=u(1+\nu)$.

$$\nu^2=(\bar{r}/p_0)^2(\delta u/h_0^2)^2.$$

Thus, the second approximation to order k_2^2 becomes

$$\left[\frac{dn_0\delta z}{dv}\right]_2 = \left(1-e_0^2\right)^{\frac{3}{2}}\left(\frac{\bar{r}}{p_0}\right)^3 \bar{W} + \left(1-e_0^2\right)^{\frac{3}{2}}\left(\frac{\bar{r}}{p_0}\right)^4\left[\left(\frac{\delta u}{h_0^2}\right)^2-\bar{W}^2\right]. \tag{8}$$

Consider the expression

$$W = \Xi + \Upsilon\cos\varphi + \Psi\sin\varphi,$$

and the definitions of Ξ, Υ, and Ψ (reference 1) we see that the next approximation after $h/h_0=h_0/h=1$ is

$$\delta\frac{h_0}{h} = -\frac{1}{3}\Xi \quad \text{and} \quad \delta\frac{h}{h_0} = +\frac{1}{3}\Xi, \tag{9}$$

whence

$$u = \bar{u} + \frac{1}{6}\Xi\bar{u} + \frac{1}{2}h_0^2\bar{W}. \tag{10}$$

After determining a first approximation of all of the perturbations to order k_2, including the λ parameters which will be discussed below, we use the results in the original expression for the disturbing function and in equation 1 to determine a second approximation of dW/dv. These terms were the k_2^2, k_3, and k_4 terms. In the process of integration some of the terms in dW/dv are divided by $(c-g)$ or a multiple of $(c-g)$. Individually, c and g are both of the order of 1, but $(c-g)$ is of the order of k_2. Hence, from the second approximation of dW/dv we obtain some first order terms in W.

The expression, to order k_2, that we obtained for Hansen's W function, is developed as follows:

$$W = \Xi + \Upsilon\cos\varphi + \Psi\sin\varphi, \tag{11}$$

where

$$\begin{aligned}
\Xi = {} & + 3k_2 h_0^4 \left(1 - 3\cos^2 i_0\right)\left[2 - \sqrt{1 - e_0^2}\right] \\
& + \frac{3}{8} k_2 h_0^4 \left(1 - \cos^2 i_0\right)\left\{- 12\cos 2\eta - 4e_0 \cos(\xi + 2\eta) - 12e_0 \cos(\xi - 2\eta)\right. \\
& \qquad \left. + \left[e_0^2 - \frac{10e_0^2 \cos^2 i_0}{1 - 5\cos^2 i_0}\right]\cos(2\xi - 2\eta)\right\} \\
& + \frac{3}{2}\frac{k_3}{k_2} h_0^2 e_0 \sin i_0 \sin(\xi - \eta) \\
& - \frac{15}{4}\frac{k_4}{k_2} h_0^4 e_0^2 \left[1 - 3\cos^2 i_0 - \frac{8\cos^4 i_0}{1 - 5\cos^2 i_0}\right]\cos(2\xi - 2\eta); \qquad (12)
\end{aligned}$$

$$\begin{aligned}
\Upsilon = {} & + k_2 h_0^4 \left(1 - 3\cos^2 i_0\right)\left[\frac{1}{2} e_0 + \frac{2e_0 + e_0^3}{1 + \sqrt{1 - e_0^2}} - \left(12 + 3e_0^2\right)\cos\xi \right. \\
& \qquad \left. - 6e_0 \cos 2\xi - e_0^2 \cos 3\xi\right] \\
& + \frac{1}{8} k_2 h_0^4 \left(1 - \cos^2 i_0\right)\left\{24e_0 \cos 2\eta + \left(28 + 5e_0^2\right)\cos(\xi + 2\eta)\right. \\
& \qquad + \left(12 - 3e_0^2\right)\cos(\xi - 2\eta) + 18e_0 \cos(2\xi + 2\eta) \\
& \qquad + \left[20e_0 + e_0^3 - \frac{\left(20e_0 + 10e_0^3\right)\cos^2 i_0}{1 - 5\cos^2 i_0}\right]\cos(2\xi - 2\eta) \\
& \qquad \left. + 3e_0^2 \cos(3\xi + 2\eta) + 3e_0^2 \cos(3\xi - 2\eta)\right\} \\
& + \frac{1}{2}\frac{k_3}{k_2} h_0^2 \left(2 + e_0^2\right)\sin i_0 \sin(\xi - \eta) \\
& - \frac{5}{4}\frac{k_4}{k_2} h_0^4 \left(2e_0 + e_0^3\right)\left[1 - 3\cos^2 i_0 - \frac{8\cos^4 i_0}{1 - 5\cos^2 i_0}\right]\cos(2\xi - 2\eta); \qquad (13)
\end{aligned}$$

$$\Psi = -\frac{1}{4} k_2 h_0^4 \left(1 - 3\cos^2 i_0\right) \left[\left(12 + 9e_0^2\right) \sin \xi + 6e_0 \sin 2\xi + e_0^2 \sin 3\xi\right]$$

$$+ \frac{1}{8} k_2 h_0^4 \left(1 - \cos^2 i_0\right) \left\{36e_0 \sin 2\eta + \left(28 + 11e_0^2\right) \sin (\xi + 2\eta)\right.$$

$$+ \left(12 - 21e_0^2\right) \sin (\xi - 2\eta) + 18e_0 \sin (2\xi + 2\eta)$$

$$+ \left[20e_0 + e_0^3 - 20e_0 \cos^2 i_0 - \frac{\left(100e_0 - 20e_0^3\right) \cos^4 i_0}{1 - 5\cos^2 i_0} + \frac{100e_0^3 \cos^6 i_0}{\left(1 - 5\cos^2 i_0\right)^2}\right] \sin (2\xi - 2\eta)$$

$$\left. + 3e_0^2 \sin (3\xi + 2\eta) + 3e_0^2 \sin (3\xi - 2\eta)\right\}$$

$$- \frac{k_3}{k_2} h_0^2 \sin i_0 \left[1 - \frac{2e_0^2 \cos^2 i_0}{1 - 5\cos^2 i_0}\right] \cos (\xi - \eta)$$

$$- \frac{5}{4} \frac{k_4}{k_2} h_0^4 \left[2e_0 + e_0^3 - \left(6e_0 + 5e_0^3\right) \cos^2 i_0\right.$$

$$\left. \frac{\left(16e_0 + 12e_0^3\right) \cos^4 i_0}{1 - 5\cos^2 i_0} + \frac{16e_0^3 \cos^6 i_0}{\left(1 - 5\cos^2 i_0\right)^2}\right] \sin (2\xi - 2\eta) . \tag{14}$$

Substituting equation 12 into equation 9 gives us the expressions for h_0/h and h/h_0 to order k_2:

$$\frac{h_0}{h} = 1 - k_2 h_0^4 \left(1 - 3\cos^2 i_0\right) \left(2 - \sqrt{1 - e_0^2}\right)$$

$$+ \frac{1}{8} k_2 h_0^4 \left(1 - \cos^2 i_0\right) \left\{12 \cos 2\eta + 4e_0 \cos (\xi + 2\eta)\right.$$

$$\left. + 12e_0 \cos (\xi - 2\eta) - \left[e_0^2 - \frac{10e_0^2 \cos^2 i_0}{1 - 5\cos^2 i_0}\right] \cos (2\xi - 2\eta)\right\}$$

$$- \frac{1}{2} \frac{k_3}{k_2} h_0^2 e_0 \sin i_0 \sin (\xi - \eta)$$

$$+ \frac{5}{4} \frac{k_4}{k_2} h_0^4 e_0^2 \left[1 - 3\cos^2 i_0 - \frac{8\cos^4 i_0}{1 - 5\cos^2 i_0}\right] \cos (2\xi - 2\eta) , \tag{15}$$

and

$$\frac{h}{h_0} = 2 - \frac{h_0}{h} .$$

The perturbations of the radius vector can be expressed most easily as the perturbations of its reciprocal u from equation 10. Remembering that $\bar{W} = W\,|_{\varphi=\xi}$ we have also

$$u = \bar{u} + \frac{1}{6} h_0^2 \left[(4 + e_0 \cos\xi)\Xi + 3\Upsilon \cos\xi + 3\Psi \sin\xi \right] . \tag{16}$$

From this we obtain

$$\begin{aligned}
u = \bar{u} &+ \frac{1}{4} k_2 h_0^6 \left(1 - 3\cos^2 i_0\right) \left[10 - 3e_0^2 - 8\sqrt{1 - e_0^2} \right. \\
&\left. + \frac{4e_0\left(1 + e_0^2\right)}{1 + \sqrt{1 - e_0^2}} \cos\xi + e_0^2 \cos 2\xi \right] \\
&- \frac{1}{32} k_2 h_0^6 \left(1 - \cos^2 i_0\right) \left\{ \left(16 + 24e_0^2\right) \cos 2\eta + 20e_0 \cos(\xi + 2\eta) \right. \\
&+ \left[8e_0 - 3e_0^3 + \frac{\left(40e_0 + 20e_0^3\right)\cos^2 i_0}{1 - 5\cos^2 i_0} - \frac{20e_0^3 \cos^4 i_0}{\left(1 - 5\cos^2 i_0\right)^2} \right] \cos(\xi - 2\eta) \\
&+ 4e_0^2 \cos(2\xi + 2\eta) - 20e_0^2 \frac{1 - 9\cos^2 i_0}{1 - 5\cos^2 i_0} \cos(2\xi - 2\eta) \\
&\left. - e_0^3 \left[1 - \frac{20\cos^2 i_0}{1 - 5\cos^2 i_0} - \frac{20\cos^4 i_0}{\left(1 - 5\cos^2 i_0\right)^2} \right] \cos(3\xi - 2\eta) \right\} \\
&- \frac{1}{4} \frac{k_3}{k_2} h_0^4 \sin i_0 \left[\left(2 + 5e_0^2\right) \sin\eta - 4e_0 \sin(\xi - \eta) + 3e_0^2 \sin(2\xi - \eta) \right] \\
&- \frac{5}{16} \frac{k_4}{k_2} h_0^6 \left(1 - 3\cos^2 i_0 - \frac{8\cos^4 i_0}{1 - 5\cos^2 i_0} \right) \left[8e_0^2 \cos(2\xi - 2\eta) \right. \\
&+ \left(e_0^3 + \frac{2e_0^3 \cos^2 i_0}{1 - 5\cos^2 i_0} \right) \cos(3\xi - 2\eta) \\
&\left. + \left(4e_0 + 3e_0^3 - \frac{2e_0^3 \cos^2 i_0}{1 - 5\cos^2 i_0} \right) \cos(\xi - 2\eta) \right] .
\end{aligned} \tag{17}$$

The final perturbation in the orbit plane to be determined is the perturbation of the mean anomaly, $n_0\,\delta z$. This is obtained by substituting the expressions for $\bar{W}$ and δu into equation 8 and integrating. It is of interest to note that any terms with an argument of $(\xi-\eta)$ or any multiple of it will, when integrated, produce a term of a lower order than in the expression for $dn_0\,\delta z/dv$. Hence, we develop $\bar{W}$ to order k_2^2 to determine all terms of order k_2 in $n_0\,\delta z$. On the other hand it might be expected that from the long period terms of order k_2 in $dn_0\,\delta z/dv$ we would obtain some terms of zero order in $n_0\,\delta z$, but this was not the case. All the long period terms generated in $dn_0\,\delta z/dv$ of order k_2 independently canceled one another. Consequently, in our expression for $n_0\,\delta z$ the short period terms come from $\bar{W}$ of order k_2, and the long period terms come from $\bar{W}$ of order k_2^2 and from the second term in equation 8, also of order k_2^2. The equation for $n_0\,\delta z$ (equation 18) is given as appendix A because of its length.

PERTURBATIONS OF THE POSITION OF THE ORBIT PLANE

The perturbations of the position of the orbit plane are developed by determining the λ parameters and the secular motions of the argument of perigee and of the ascending node. By definition, the osculating values of the argument of perigee and of the node are

$$\omega = (g - c)v + \pi_0 - \sigma_0 + N + K + \phi \qquad (19)$$

and

$$\theta = (1 - h')v + \theta_0 - N + K. \qquad (20)$$

The λ parameters are defined by

$$\begin{aligned} \lambda_1 &= \sin\tfrac{i}{2}\cos N, \\ \lambda_2 &= \sin\tfrac{i}{2}\sin N, \\ \lambda_3 &= \cos\tfrac{i}{2}\sin K, \\ \lambda_4 &= \cos\tfrac{i}{2}\cos K. \end{aligned} \qquad (21)$$

They include all the periodic perturbations from the position of the mean node, from the mean argument of perigee, and from the inclination. The derivatives of the λ parameters as given by Musen are

$$\frac{d\lambda_1}{dv} = +\left(\frac{h' + g}{2} - 1\right)\lambda_2 + \frac{h_0^2}{4}\left(\frac{h}{h_0}\right)^2 \frac{1}{u^2}\frac{\partial\Omega}{\partial\lambda_2}\cos i, \qquad (22)$$

$$\frac{d\lambda_2}{dv} = -\left(\frac{h' + g}{2} - 1\right)\lambda_1 - \frac{h_0^2}{4}\left(\frac{h}{h_0}\right)^2 \frac{1}{u^2}\frac{\partial\Omega}{\partial\lambda_1}\cos i, \qquad (23)$$

$$\frac{d\lambda_3}{dv} = +\frac{h' - g}{2}\lambda_4 + \frac{h_0^2}{4}\left(\frac{h}{h_0}\right)^2 \frac{1}{u^2}\frac{\partial\Omega}{\partial\lambda_4}\cos i, \qquad (24)$$

$$\frac{d\lambda_4}{dv} = -\frac{h' - g}{2}\lambda_3 - \frac{h_0^2}{4}\left(\frac{h}{h_0}\right)^2 \frac{1}{u^2}\frac{\partial\Omega}{\partial\lambda_3}\cos i. \qquad (25)$$

By determining the derivatives of Ω with respect to the λ's to order k_2 and substituting them into the foregoing equations, we develop a series for each of the λ parameters. The secular terms h' and g are determined from equations 23 and 24 in such a manner that no constant term should appear on the right-hand sides. And from the second approximations of the λ derivatives, we determine these secular terms

to order k_2^2, that is,

$$g = 1 + 3k_2 h_0^4 \cos^2 i_0 - \frac{15}{4} k_4 h_0^8 \left(2 + 3e_0^2\right)\left(3\cos^2 i_0 - 7\cos^4 i_0\right) + \frac{3}{8} k_2^2 h_0^8 \left[\left(110 - 48\sqrt{1-e_0^2} + 13e_0^2\right)\cos^2 i_0 - \left(290 - 144\sqrt{1-e_0^2} + 9e_0^2\right)\cos^4 i_0\right]; \tag{26}$$

$$(1 - h') = -3k_2 h_0^4 \cos i_0 + \frac{15}{4} k_4 h_0^8 \left(2 + 3e_0^2\right)\left(3\cos i_0 - 7\cos^3 i_0\right) - \frac{3}{8} k_2^2 h_0^8 \left[\left(88 - 40\sqrt{1-e_0^2} + 9e_0^2\right)\cos i_0 - \left(236 - 120\sqrt{1-e_0^2} + 5e_0^2\right)\cos^3 i_0\right]. \tag{27}$$

From equations 26 and 2 we obtain the secular motion of the argument of perigee:

$$g - c = -\frac{3}{2} k_2 h_0^4 \left(1 - 5\cos^2 i_0\right) + \frac{15}{16} k_4 h_0^8 \left[\left(12 + 9e_0^2\right) - \left(144 + 126e_0^2\right)\cos^2 i_0 + \left(196 + 189e_0^2\right)\cos^4 i_0\right] - \frac{3}{32} k_2^2 h_0^8 \left[\left(162 - 64\sqrt{1-e_0^2} + 25e_0^2\right) - \left(1500 - 672\sqrt{1-e_0^2} + 126e_0^2\right)\cos^2 i_0 + \left(2810 - 1440\sqrt{1-e_0^2} + 45e_0^2\right)\cos^4 i_0\right]. \tag{28}$$

The analytic expressions for the λ parameters obtained are as follows:

$$\lambda_1 = \sin\frac{i_0}{2} + k_2 h_0^4 \sin i_0 \cos i_0 \cos\frac{i_0}{2}\left[\frac{3}{4}\cos 2\eta + \frac{1}{4} e_0 \cos(\xi + 2\eta) + \frac{3}{4} e_0 \cos(\xi - 2\eta) - \frac{e_0^2\left(1 - 15\cos^2 i_0\right)}{16\left(1 - 5\cos^2 i_0\right)}\cos(2\xi - 2\eta)\right] + -\frac{1}{4}\frac{k_3}{k_2} h_0^2 e_0 \cos i_0 \cos\frac{i_0}{2}\sin(\xi - \eta) + \frac{5}{8}\frac{k_4}{k_2} h_0^4 e_0^2 \sin i_0 \cos i_0 \cos\frac{i_0}{2}\frac{\left(1 - 7\cos^2 i_0\right)}{\left(1 - 5\cos^2 i_0\right)}\cos(2\xi - 2\eta); \tag{29}$$

$$\lambda_2 = + k_2 h_0^4 \sin i_0 \cos i_0 \cos \frac{i_0}{2} \left\{ \frac{3}{2} e_0 \sin \xi - \frac{3}{4} \sin 2\eta - \frac{1}{4} e_0 \sin (\xi + 2\eta) \right.$$

$$+ \frac{3}{4} e_0 \sin (\xi - 2\eta) + \left[- \frac{e_0^2 \left(8 + 3 \cos i_0 - 18 \cos^2 i_0\right)}{8\left(1 - 5 \cos^2 i_0\right)} \right.$$

$$\left. \left. + \frac{e_0^2 \left(5 + \cos i_0 - 6 \cos^2 i_0\right)\left(1 - 15 \cos^2 i_0\right)}{16 \left(1 - 5 \cos^2 i_0\right)^2} \right] \sin (2\xi - 2\eta) \right\}$$

$$+ \frac{1}{4} \frac{k_3}{k_2} h_0^2 e_0 \cos i_0 \cos \frac{i_0}{2} \left(\frac{1 - 2 \cos i_0 - 3 \cos^2 i_0}{1 - 5 \cos^2 i_0} \right) \cos (\xi - \eta)$$

$$+ \frac{5}{4} \frac{k_4}{k_2} h_0^4 e_0^2 \sin i_0 \cos i_0 \cos \frac{i_0}{2} \left[\frac{4 - 7 \cos^2 i_0}{1 - 5 \cos^2 i_0} \right.$$

$$\left. - \frac{\left(5 + \cos i_0 - 6 \cos^2 i_0\right)\left(1 - 7 \cos^2 i_0\right)}{2 \left(1 - 5 \cos^2 i_0\right)^2} \right] \sin (2\xi - 2\eta) \quad . \tag{30}$$

$$\lambda_3 = - k_2 h_0^4 \sin i_0 \cos i_0 \sin \frac{i_0}{2} \left\{ \frac{3}{2} e_0 \sin \xi - \frac{3}{4} \sin 2\eta - \frac{1}{4} e_0 \sin (\xi + 2\eta) \right.$$

$$+ \frac{3}{4} e_0 \sin (\xi - 2\eta) + \left[- \frac{e_0^2 \left(8 - 3 \cos i_0 - 18 \cos^2 i_0\right)}{8\left(1 - 5 \cos^2 i_0\right)} \right.$$

$$\left. \left. + \frac{e_0^2 \left(5 - \cos i_0 - 6 \cos^2 i_0\right)\left(1 - 15 \cos^2 i_0\right)}{16 \left(1 - 5 \cos^2 i_0\right)^2} \right] \sin (2\xi - 2\eta) \right\}$$

$$- \frac{1}{4} \frac{k_3}{k_2} h_0^2 e_0 \cos i_0 \sin \frac{i_0}{2} \left(\frac{1 + 2 \cos i_0 - 3 \cos^2 i_0}{1 - 5 \cos^2 i_0} \right) \cos (\xi - \eta)$$

$$- \frac{5}{4} \frac{k_4}{k_2} h_0^4 e_0^2 \sin i_0 \cos i_0 \sin \frac{i_0}{2} \left[\frac{4 - 7 \cos^2 i_0}{1 - 5 \cos^2 i_0} - \frac{\left(5 - \cos i_0 - 6 \cos^2 i_0\right)\left(1 - 7 \cos^2 i_0\right)}{2\left(1 - 5 \cos^2 i_0\right)^2} \right] \sin (2\xi - 2\eta) \quad ; \tag{31}$$

$$\lambda_4 = \cos\frac{i_0}{2} - k_2 h_0^4 \sin i_0 \cos i_0 \sin\frac{i_0}{2}\left[\frac{3}{4}\cos 2\eta + \frac{1}{4} e_0 \cos(\xi + 2\eta) + \frac{3}{4} e_0 \cos(\xi - 2\eta) - \frac{1}{16} e_0^2 \frac{1 - 15\cos^2 i_0}{1 - 5\cos^2 i_0}\cos(2\xi - 2\eta)\right]$$

$$+ \frac{1}{4}\frac{k_3}{k_2} h_0^2 e_0 \cos i_0 \sin\frac{i_0}{2} \sin(\xi - \eta)$$

$$- \frac{5}{8}\frac{k_4}{k_2} h_0^4 e_0^2 \sin i_0 \cos i_0 \sin\frac{i_0}{2}\left(\frac{1 - 7\cos^2 i_0}{1 - 5\cos^2 i_0}\right)\cos(2\xi - 2\eta) \cdot \qquad (32)$$

CONCLUSION

An analytic development of Musen's true anomaly version of the Hansen theory to the first order has been presented. Expressions for the W function, u, h/h_0, $n_0\,\delta z$, and the λ-parameeters have been given, as well as analytic representations of the secular motions of the node and argument of perigee for this theory.

For nearly circular orbits the expression for $(1-c)$ can be inserted directly into the numerical program of this theory, thus eliminating the only division by e_0 which could be a point of slow or no convergence.

We see that this theory, as well as other general perturbation theories,[2,3,4] cannot be used for orbits with inclinations in the neighborhood of 63.4 degrees: since that is the value for which $(1-5\cos^2 i_0)$ goes to zero, the series for the perturbations will not converge and will have no meaning.

The appearance of a coupling term k_3k_4/k_2^3 in the perturbations of the mean anomaly was an unexpected result. This is due to the use of the pseudo time in Hansen's theory.

Finally we conclude that, whereas an analytic development is useful for understanding the makeup of this theory, for comparison with other general perturbation theories, and for checking out the coding of the numerical program, it is impracticable for general use even to the first order.

ACKNOWLEDGMENTS

The authors wish to express their appreciation to Dr. P. Musen for several valuable discussions concerning this report and to Mrs. Aileen Marlow who verified the mathematical results.

REFERENCES

1. Musen, P., "The Theory of Artificial Satellites in Terms of the Orbital True Longitude," *J. Geophys. Res.*, **66**(2): 403–409, February 1961.
2. Brouwer, D., "Solution of the Problem of Artificial Satellite Theory without Drag," *Astronom. J.*, **64**(9): 378–397, November 1959.
3. Kozai, Y., "The Motion of a Close Earth Satellite," *Astronom. J.*, **64**(9): 367–377, November 1959.
4. Garfinkel, B., "The Orbit of a Satellite of an Oblate Planet," *Astronom. J.*, **64**(9): 353–367, November 1959.

Appendix A

EQUATION FOR THE PERTURBATION OF THE MEAN ANOMALY

Because of its length equation 18, the perturbation of the mean anomaly, was not included in the text and is given here.

To reduce the length of equation 18, the following substitutions are used:

$$C_0' = 4 - 2\sqrt{1 - e_0^2}\,.$$

$$C_1' = 1 - \frac{2}{3}\sqrt{1 - e_0^2} + \frac{4}{3}\left(1 + \sqrt{1 - e_0^2}\right)^{-1}.$$

$$\beta = \frac{1 - \sqrt{1 - e_0^2}}{e_0}.$$

$$A = (1 + \beta^2)(10 - 5\beta^2 + \beta^4)\,.$$

$$B = (1 + \beta^2)^2(5 - 4\beta^2 + \beta^4)\,.$$

$$D = (1 + \beta^2)^3(7 - 7\beta^2 + 2\beta^4)\,.$$

$$E = (1 + \beta^2)^4(14 - 16\beta^2 + 5\beta^4)\,.$$

$$F = (1 + \beta^2)^5(12 - 15\beta^2 + 5\beta^4)\,.$$

$$G = (1 + \beta^2)(35 - 21\beta^2 + 7\beta^4 - \beta^6)\,.$$

$$H = (1 + \beta^2)^2(14 - 14\beta^2 + 6\beta^4 - \beta^6)\,.$$

$$I = (1 + \beta^2)^3(42 - 54\beta^2 + 27\beta^4 - 5\beta^6).$$

With these substitutions equation 18 is:

$$\begin{aligned} n_0\delta z = \frac{k_2 h_0^4 (1 - 3\cos^2 i_0)}{(1 - e_0^2)} \Bigg\{ & \left[e_0^2\left(-\frac{25}{8} + \frac{9}{4}C_0' - \frac{9}{8}C_1'\right) - \frac{17}{8}e_0^4 + \frac{3}{32}e_0^4(1 - C_1')A + \frac{3}{64}e_0^4 B\right]\sin 2\xi \\ & + \left[e_0^3\left(-1 + \frac{3}{4}C_1'\right) + \frac{e_0^3}{8}(2 - C_0' + e_0^2)A - \frac{3}{32}e_0^5(1 - C_1')B - \frac{e_0^7}{64}D\right]\sin 3\xi \\ & + \left[\frac{3}{16}e_0^4 + \frac{3}{64}e_0^4(1 - C_1')A - \frac{9}{64}e_0^4(2 - C_0' + e_0^2)B + \frac{9}{256}e_0^6(1 - C_1')D + \frac{e_0^8}{256}E\right]\sin 4\xi \\ & + \left[-\frac{e_0^5}{80}A - \frac{9}{160}e_0^5(1 - C_1')B + \frac{9}{160}e_0^5(2 - C_0' + e_0^2)D - \frac{3}{320}e_0^7(1 - C_1')E\right]\sin 5\xi \\ & + \left[\frac{e_0^6}{64}B + \frac{3}{128}e_0^6(1 - C_1')D - \frac{e_0^6}{64}(2 - C_0' + e_0^2)E + \frac{3}{512}e_0^8(1 - C_1')F\right]\sin 6\xi \Bigg\} \end{aligned}$$

$$+\frac{k_2 h_0^4 (1-\cos^2 i_0)}{(1-e_0^2)}\left\{\left(\frac{1}{4}+\frac{7}{16}e_0^2+e_0^4\right)\sin 2\eta+\left(\frac{3}{2}e_0+\frac{15}{8}e_0^3\right)\sin(\xi-2\eta)\right.$$
$$+\left[-\frac{e_0}{4}-\frac{e_0^3}{8}-\frac{e_0^5}{16}A\right]\sin(\xi+2\eta)+\left[\frac{3}{16}e_0^2-\frac{3}{16}e_0^4+\frac{3}{128}e_0^4 A+\frac{9}{128}e_0^6 B\right]\sin(2\xi+2\eta)$$
$$+\left[-\frac{27}{8}e_0^3-\frac{e_0^5}{16}(1-e_0^2)A\right]\sin(3\xi-2\eta)+\left[-\frac{e_0^5}{80}(1-e_0^2)A-\frac{9}{320}e_0^5 B-\frac{9}{320}e_0^7 D\right]\sin(3\xi+2\eta)$$
$$+\left[\frac{9}{8}e_0^4+\frac{3}{64}e_0^4 A+\frac{3}{64}e_0^4(1-e_0^2)B\right]\sin(4\xi-2\eta)+\left[\frac{e_0^4}{64}(1-e_0^2)B+\frac{3}{256}e_0^6 D+\frac{e_0^8}{128}E\right]\sin(4\xi+2\eta)$$
$$+\left[-\frac{e_0^5}{16}A-\frac{3}{64}e_0^5 B-\frac{e_0^5}{64}(1-e_0^2)D\right]\sin(5\xi-2\eta)+\left[-\frac{3}{448}e_0^5(1-e_0^2)D-\frac{3}{896}e_0^7 E\right]\sin(5\xi+2\eta)$$
$$\left.+\left[\frac{9}{128}e_0^6 B+\frac{9}{512}e_0^6 D+\frac{e_0^6}{256}(1-e_0^2)E\right]\sin(6\xi-2\eta)+\left[\frac{e_0^6}{512}(1-e_0^2)E\right]\sin(6\xi+2\eta).\right\}$$

$$+\frac{k_2 h_0^4}{(1-e_0^2)(1-5\cos^2 i_0)}\left\{\left[-\frac{15}{8}e_0^2+\frac{e_0^6}{64}A\right]\sin 2\eta+\left(-\frac{5}{2}e_0-\frac{17}{16}e_0^3+\frac{3}{16}e_0^5\right)\sin(\xi-2\eta)\right.$$
$$+\left[\frac{5}{4}e_0^2+\frac{3}{16}e_0^3-\frac{e_0^5}{64}A-\frac{e_0^7}{64}B\right]\sin(\xi+2\eta)+\left[\left(-\frac{5}{64}e_0^4-\frac{3}{256}e_0^6\right)A+\frac{9}{512}e_0^6 B+\frac{3}{512}e_0^8 D\right]\sin(2\xi+2\eta)$$
$$+\left(\frac{47}{16}e_0^3+\frac{7}{16}e_0^5\right)\sin(3\xi-2\eta)+\left[\left(\frac{3}{32}e_0^5+\frac{9}{640}e_0^7\right)B-\frac{9}{1280}e_0^7 D\right]\sin(3\xi+2\eta)$$
$$+\left[\frac{15}{32}e_0^4-\left(\frac{5}{32}e_0^4+\frac{3}{128}e_0^6\right)A\right]\sin(4\xi-2\eta)+\left[\left(-\frac{5}{128}e_0^6-\frac{3}{512}e_0^8\right)D+\frac{e_0^8}{512}E\right]\sin(4\xi+2\eta)$$
$$+\left[-\frac{e_0^5}{8}-\frac{e_0^5}{64}A+\left(\frac{5}{32}e_0^5+\frac{3}{128}e_0^7\right)B\right]\sin(5\xi-2\eta)+\left[\frac{5}{448}e_0^7 E\right]\sin(5\xi+2\eta)$$
$$\left.+\left[\frac{e_0^6}{128}A+\frac{9}{512}e_0^6 B+\left(-\frac{15}{256}e_0^6-\frac{9}{1024}e_0^8\right)D\right]\sin(6\xi-2\eta)+\left[-\frac{15}{256}e_0^8 F\right]\sin(6\xi+2\eta)\right\}$$

$$+\frac{k_2 h_0^4 \cos^2 i_0}{(1-e_0^2)(1-5\cos^2 i_0)}\left\{\left[\frac{105}{8}e_0^2-\frac{9}{16}e_0^4-\frac{13}{64}e_0^6 A\right]\sin 2\eta+\left(\frac{35}{2}e_0+5e_0^3-\frac{9}{4}e_0^5\right)\sin(\xi-2\eta)\right.$$
$$+\left[-\frac{35}{4}e_0^3-\frac{21}{8}e_0^5+\frac{e_0^5}{4}A+\frac{13}{64}e_0^7 B\right]\sin(\xi+2\eta)+\left[\left(\frac{35}{64}e_0^4+\frac{21}{128}e_0^6\right)A-\frac{9}{32}e_0^6 B-\frac{39}{512}e_0^8 D\right]\sin(2\xi+2\eta)$$
$$+\left(-14e_0^3-\frac{25}{4}e_0^5\right)\sin(3\xi-2\eta)+\left[\left(-\frac{21}{32}e_0^5-\frac{63}{320}e_0^7\right)B+\frac{9}{80}e_0^7 D\right]\sin(3\xi+2\eta)+\left[-\frac{111}{16}e_0^4+\left(\frac{35}{32}e_0^4+\frac{21}{64}e_0^6\right)A\right]\sin(4\xi-2\eta)$$
$$+\left[\left(\frac{35}{128}e_0^6+\frac{21}{256}e_0^8\right)D-\frac{e_0^8}{32}E\right]\sin(4\xi+2\eta)+\left[\frac{13}{8}e_0^5+\frac{e_0^5}{4}A+\left(-\frac{35}{32}e_0^5-\frac{21}{64}e_0^7\right)B\right]\sin(5\xi-2\eta)$$
$$\left.+\left[-\frac{5}{64}e_0^7 E\right]\sin(5\xi+2\eta)+\left[-\frac{13}{128}e_0^6 A-\frac{9}{32}e_0^6 B+\left(\frac{105}{256}e_0^6+\frac{63}{512}e_0^8\right)D\right]\sin(6\xi-2\eta)+\left[\frac{105}{256}e_0^8 F\right]\sin(6\xi+2\eta)\right\}$$

$$+\frac{k_2 h_0^4 \cos^4 i_0}{(1-e_0^2)(1-5\cos^2 i_0)}\left\{\left[-\frac{45}{4}e_0^2+\frac{9}{16}e_0^4+\frac{3}{16}e_0^6 A\right]\sin 2\eta+\left(-15e_0-\frac{63}{16}e_0^3+\frac{33}{16}e_0^5\right)\sin(\xi-2\eta)\right.$$
$$+\left[\frac{15}{2}e_0^3+\frac{39}{16}e_0^5-\frac{15}{64}e_0^5 A-\frac{3}{16}e_0^7 B\right]\sin(\xi+2\eta)+\left[\left(-\frac{15}{32}e_0^4-\frac{39}{256}e_0^6\right)A+\frac{135}{512}e_0^6 B+\frac{9}{128}e_0^8 D\right]\sin(2\xi+2\eta)$$
$$+\left(\frac{177}{16}e_0^3+\frac{93}{16}e_0^5\right)\sin(3\xi-2\eta)+\left[\left(\frac{9}{16}e_0^5+\frac{117}{640}e_0^7\right)B-\frac{27}{256}e_0^7 D\right]\sin(3\xi+2\eta)$$
$$+\left[\frac{207}{32}e_0^4+\left(-\frac{15}{16}e_0^4-\frac{39}{128}e_0^6\right)A\right]\sin(4\xi-2\eta)+\left[\left(-\frac{15}{64}e_0^6-\frac{39}{512}e_0^8\right)D+\frac{15}{512}e_0^8 E\right]\sin(4\xi+2\eta)$$
$$+\left[-\frac{3}{2}e_0^5-\frac{15}{64}e_0^5 A+\left(\frac{15}{16}e_0^5+\frac{39}{128}e_0^7\right)B\right]\sin(5\xi-2\eta)+\left[\frac{15}{224}e_0^7 E\right]\sin(5\xi+2\eta)$$
$$\left.+\left[\frac{3}{32}e_0^6 A+\frac{135}{512}e_0^6 B+\left(-\frac{45}{128}e_0^6-\frac{117}{1024}e_0^8\right)D\right]\sin(6\xi-2\eta)+\left[-\frac{45}{128}e_0^8 F\right]\sin(6\xi+2\eta)\right\}$$

$$+\frac{k_3 h_0^2 \sin i_0}{k_2(1-e_0^2)}\left\{\left(1+\frac{e_0^2}{2}-\frac{3}{2}e_0^4\right)\cos\eta+\left[-\frac{3}{4}e_0-\frac{9}{16}e_0^3+\frac{7}{64}e_0^5 A\right]\cos(\xi+\eta)\right.$$
$$+\frac{5}{2}e_0^2(1-e_0^2)\cos(2\xi-\eta)+\left[\frac{e_0^2}{2}\left(1+\frac{9}{4}e_0^2\right)-\frac{e_0^4}{16}A-\frac{7}{64}e_0^6 B\right]\cos(2\xi+\eta)$$
$$+\left[-\frac{39}{16}e_0^3+\frac{e_0^3}{16}\left(1+\frac{9}{4}e_0^2\right)A\right]\cos(3\xi-\eta)+\left[-\frac{e_0^3}{32}\left(1+\frac{9}{4}e_0^2\right)A+\frac{9}{128}e_0^5 B+\frac{21}{512}e_0^7 D\right]\cos(3\xi+\eta)$$
$$+\left[\frac{7}{8}e_0^4+\frac{e_0^4}{16}A-\frac{e_0^4}{16}\left(1+\frac{9}{4}e_0^2\right)B\right]\cos(4\xi-\eta)+\left[\frac{3}{80}e_0^4\left(1+\frac{9}{4}e_0^2\right)B-\frac{9}{320}e_0^6 D-\frac{7}{640}e_0^8 E\right]\cos(4\xi+\eta)$$
$$+\left[-\frac{7}{128}e_0^5 A-\frac{9}{128}e_0^5 B+\frac{3}{128}e_0^5\left(1+\frac{9}{4}e_0^2\right)D\right]\cos(5\xi-\eta)+\left[-\frac{e_0^5}{64}\left(1+\frac{9}{4}e_0^2\right)D+\frac{e_0^7}{128}E\right]\cos(5\xi+\eta)$$
$$\left.+\left[\frac{21}{320}e_0^6 B+\frac{9}{320}e_0^6 D-\frac{e_0^6}{160}\left(1+\frac{9}{4}e_0^2\right)E\right]\cos(6\xi-\eta)+\left[\frac{e_0^6}{224}\left(1+\frac{9}{4}e_0^2\right)E\right]\cos(6\xi+\eta)\right\}$$

$$+\frac{k_4 h_0^4(1-8\cos^2 i_0+7\cos^4 i_0)}{k_2(1-e_0^2)(1-5\cos^2 i_0)}\left\{\left[\frac{15}{8}e_0^2-\frac{5}{32}e_0^6 A\right]\sin 2\eta+\left(\frac{5}{2}e_0-\frac{5}{8}e_0^3-\frac{15}{8}e_0^5\right)\sin(\xi-2\eta)\right.$$
$$+\left[-\frac{5}{4}e_0^3\left(1+\frac{3}{2}e_0^2\right)+\frac{5}{32}e_0^5 A+\frac{5}{32}e_0^7 B\right]\sin(\xi+2\eta)+\left[\frac{5}{64}e_0^4\left(1+\frac{3}{2}e_0^2\right)A-\frac{45}{256}e_0^6 B-\frac{15}{256}e_0^8 D\right]\sin(2\xi+2\eta)$$
$$+\frac{35}{8}e_0^3(1-e_0^2)\sin(3\xi-2\eta)+\left[-\frac{3}{32}e_0^5\left(1+\frac{3}{2}e_0^2\right)B+\frac{9}{128}e_0^7 D\right]\sin(3\xi+2\eta)$$
$$+\left[-\frac{75}{16}e_0^4+\frac{5}{32}e_0^4\left(1+\frac{3}{2}e_0^2\right)A\right]\sin(4\xi-2\eta)+\left[\frac{5}{128}e_0^6\left(1+\frac{3}{2}e_0^2\right)D-\frac{5}{256}e_0^8 E\right]\sin(4\xi+2\eta)$$
$$+\left[\frac{5}{4}e_0^5+\frac{5}{32}e_0^5 A-\frac{5}{32}e_0^5\left(1+\frac{3}{2}e_0^2\right)B\right]\sin(5\xi-2\eta)+\left[-\frac{5}{448}e_0^7 E\right]\sin(5\xi+2\eta)$$
$$\left.+\left[-\frac{5}{64}e_0^6 A-\frac{45}{256}e_0^6 B+\frac{15}{256}e_0^6\left(1+\frac{3}{2}e_0^2\right)D\right]\sin(6\xi-2\eta)\right\}$$

$$+\frac{k_2 h_0^4(2-41\cos^2 i_0+174\cos^4 i_0-135\cos^6 i_0)}{(1-e_0^2)(1-5\cos^2 i_0)^2}\left\{\left[\frac{3}{32}e_0^4-\frac{1}{128}e_0^6 A\right]\sin 2\eta+\frac{e_0^3}{8}(1-e_0^2)\sin(\xi-2\eta)\right.$$
$$+\left[-\frac{1}{16}e_0^5+\frac{1}{128}e_0^7 B\right]\sin(\xi+2\eta)+\left[\frac{1}{256}e_0^6 A-\frac{3}{1024}e_0^8 D\right]\sin(2\xi+2\eta)+\frac{e_0^3}{8}(1-e_0^2)\sin(3\xi-2\eta)-\frac{3}{640}e_0^7 B\sin(3\xi+2\eta)$$
$$\left.+\left[-\frac{3}{32}e_0^4+\frac{1}{128}e_0^6 A\right]\sin(4\xi-2\eta)+\frac{1}{512}e_0^8 D\sin(4\xi+2\eta)+\left[\frac{1}{16}e_0^5-\frac{1}{128}e_0^7 B\right]\sin(5\xi-2\eta)+\left[-\frac{1}{256}e_0^6 A+\frac{3}{1024}e_0^8 D\right]\sin(6\xi-2\eta)\right\}$$

$$+\frac{k_3 h_0^3 \sin i_0(2-9\cos^2 i_0)}{k_2(1-e_0^2)(1-5\cos^2 i_0)}\left\{\left(-e_0^2+e_0^4\right)\cos\eta+\left[\frac{3}{4}e_0^3-\frac{1}{16}e_0^5 A\right]\cos(\xi+\eta)+\left(-e_0^3+e_0^4\right)\cos(2\xi-\eta)+\left[-\frac{1}{2}e_0^4+\frac{1}{16}e_0^6 B\right]\cos(2\xi+\eta)\right.$$
$$+\left[\frac{3}{4}e_0^3-\frac{1}{16}e_0^5 A\right]\cos(3\xi-\eta)+\left[\frac{1}{32}e_0^5 A-\frac{3}{128}e_0^7 D\right]\cos(3\xi+\eta)+\left[-\frac{1}{2}e_0^4+\frac{1}{16}e_0^6 B\right]\cos(4\xi-\eta)+\left[-\frac{3}{80}e_0^6 B+\frac{1}{160}e_0^8 E\right]\cos(4\xi+\eta)$$
$$\left.+\left[\frac{1}{32}e_0^5 A-\frac{3}{128}e_0^7 D\right]\cos(5\xi-\eta)+\frac{1}{64}e_0^7 D\cos(5\xi+\eta)+\left[-\frac{3}{80}e_0^6 B+\frac{1}{160}e_0^8 E\right]\cos(6\xi-\eta)-\frac{1}{224}e_0^8 E\cos(6\xi+\eta)\right\}$$

$$+\frac{k_4 h_0^4(2-25\cos^2 i_0+86\cos^4 i_0-63\cos^6 i_0)}{k_2(1-e_0^2)(1-5\cos^2 i_0)^2}\left\{\left[-\frac{15}{16}e_0^4+\frac{5}{64}e_0^6 A\right]\sin 2\eta-\frac{5}{4}e_0^3(1-e_0^2)\sin(\xi-2\eta)+\left[\frac{5}{8}e_0^5-\frac{5}{64}e_0^7 B\right]\sin(\xi+2\eta)\right.$$
$$+\left[-\frac{5}{128}e_0^6 A+\frac{15}{512}e_0^8 D\right]\sin(2\xi+2\eta)-\frac{5}{4}e_0^3(1-e_0^2)\sin(3\xi-2\eta)+\frac{3}{64}e_0^7 B\sin(3\xi+2\eta)+\left[\frac{15}{16}e_0^4-\frac{5}{64}e_0^6 A\right]\sin(4\xi-2\eta)$$
$$\left.-\frac{5}{256}e_0^8 D\sin(4\xi+2\eta)+\left[-\frac{5}{8}e_0^5+\frac{5}{64}e_0^7 B\right]\sin(5\xi-2\eta)+\left[\frac{5}{128}e_0^6 A-\frac{15}{512}e_0^8 D\right]\sin(6\xi-2\eta)\right\}$$

$$+\frac{k_2 h_0^4}{(1-e_0^2)(1-5\cos^2 i_0)}\left[-\frac{15}{16}e_0^2-\frac{23}{32}e_0^4+\frac{7}{4}e_0^2C_0'+\frac{5}{32}e_0^4C_0'+\frac{15}{16}e_0^2C_1'+\frac{15}{32}e_0^4C_1'\right.$$
$$\left.+\left(\frac{9}{128}e_0^4+\frac{e_0^6}{32}-\frac{11}{384}e_0^4C_0'-\frac{3}{128}e_0^4C_1'\right)A+\left(-\frac{31}{2560}e_0^4+\frac{1}{256}e_0^6C_0'+\frac{3}{256}e_0^6C_1'\right)B-\frac{e_0^8}{1024}D\right]\sin(2\xi-2\eta)$$

$$+\frac{k_2 h_0^4\cos^2 i_0}{(1-e_0^2)(1-5\cos^2 i_0)}\left[-\frac{75}{8}e_0^2+\frac{31}{4}e_0^4-7e_0^2C_0'-\frac{5}{8}e_0^4C_0'-\frac{15}{4}e_0^2C_1'-\frac{15}{8}e_0^4C_1'\right.$$
$$\left.+\left(-\frac{33}{64}e_0^4-\frac{25}{128}e_0^6+\frac{11}{96}e_0^4C_0'+\frac{3}{32}e_0^4C_1'\right)A+\left(\frac{47}{320}e_0^4-\frac{1}{64}e_0^6C_0'-\frac{3}{64}e_0^6C_1'\right)B-\frac{e_0^8}{512}D\right]\sin(2\xi-2\eta)$$

$$+\frac{k_2 h_0^4\cos^4 i_0}{(1-e_0^2)(1-5\cos^2 i_0)}\left[\frac{165}{16}e_0^2-\frac{225}{32}e_0^4+\frac{21}{4}e_0^2C_0'+\frac{15}{32}e_0^4C_0'+\frac{45}{16}e_0^2C_1'+\frac{45}{32}e_0^4C_1'\right.$$
$$\left.+\left(\frac{57}{128}e_0^4+\frac{21}{128}e_0^6-\frac{11}{128}e_0^4C_0'-\frac{9}{128}e_0^4C_1'\right)A+\left(-\frac{69}{512}e_0^4+\frac{3}{256}e_0^6C_0'+\frac{9}{256}e_0^6C_1'\right)B+\frac{3}{1024}e_0^8D\right]\sin(2\xi-2\eta)$$

$$+\frac{k_2 h_0^4(1-2\cos^2 i_0+\cos^4 i_0)}{(1-e_0^2)(1-5\cos^2 i_0)}\left[-\frac{75}{256}e_0^4+\left(\frac{13}{256}e_0^4+\frac{5}{1024}e_0^6\right)A+\left(-\frac{3}{128}e_0^4-\frac{3}{64}e_0^6\right)B+\left(\frac{13}{1024}e_0^6+\frac{9}{2048}e_0^8\right)D-\frac{5}{3072}e_0^8E\right]\sin(4\xi-4\eta)$$

$$+\frac{k_4 h_0^4(1-8\cos^2 i_0+7\cos^4 i_0)}{k_2(1-e_0^2)(1-5\cos^2 i_0)}\left[-\frac{45}{4}e_0^2+\frac{25}{8}e_0^4+\frac{5}{16}e_0^6-\frac{5}{64}e_0^4(1+e_0^2)A+\frac{e_0^6}{128}\left(\frac{9}{2}+e_0^2\right)B-\frac{1}{512}e_0^8D\right]\sin(2\xi-2\eta)$$

$$+\frac{k_4 h_0^4(1-2\cos^2 i_0+\cos^4 i_0)}{k_2(1-e_0^2)(1-5\cos^2 i_0)}\left[\frac{105}{64}e_0^4+\frac{35}{256}e_0^6-\frac{35}{256}e_0^4\left(1+\frac{e_0^2}{2}\right)A+\frac{3}{256}e_0^4\left(7+\frac{35}{2}e_0^2+\frac{7}{4}e_0^4\right)B-\frac{21}{512}e_0^6\left(\frac{3}{2}+e_0^2\right)D+\frac{13}{1024}e_0^8E\right]\sin(4\xi-4\eta)$$

$$+\frac{k_3 h_0^2\sin i_0}{k_2(1-e_0^2)}\left[-\frac{23}{4}e_0+\frac{11}{16}e_0^3+\frac{5}{32}e_0^5-\frac{e_0^5}{32}A+\frac{e_0^7}{256}B\right]\cos(\xi-\eta)$$

$$+\frac{k_3 h_0^2\sin i_0(1-\cos^2 i_0)}{k_2(1-e_0^2)(1-5\cos^2 i_0)}\left[\frac{85}{144}e_0^3+\frac{5}{144}e_0^5-\frac{5}{288}e_0^3\left(1+\frac{3}{4}e_0^2\right)A+\frac{e_0^5}{128}\left(3+\frac{e_0^2}{2}\right)B-\frac{7}{1536}e_0^7D\right]\cos(3\xi-3\eta)$$

$$+\frac{k_2 h_0^4(1-4\cos^2 i_0+3\cos^4 i_0)}{(1-e_0^2)^2(1-5\cos^2 i_0)}\left[\frac{e_0^2}{2}\left(11-\frac{23}{4}C_0'+\frac{3}{2}C_1'\right)+\frac{e_0^4}{192}(315-233C_0'+621C_1')-\frac{e_0^6}{192}(141-35C_0'-105C_1')\right.$$
$$\left.+\frac{e_0^6}{1536}(15-13C_0'-15C_1'+8e_0^2)G+\frac{e_0^8}{384}\left(1+\frac{1}{2}C_0'+\frac{3}{2}C_1'\right)H\right]\sin(2\xi-2\eta)$$

$$+\frac{k_2 h_0^4}{(1-e_0^2)^2(1-5\cos^2 i_0)^2}\left[e_0^2\left(-\frac{97}{16}+\frac{155}{48}C_0'-\frac{15}{16}C_1'\right)+e_0^4\left(-\frac{31}{32}+\frac{91}{96}C_0'-\frac{105}{32}C_1'\right)+\frac{1}{24}e_0^6C_0'\right]\sin(2\xi-2\eta)$$

$$+\frac{k_2 h_0^4\cos^2 i_0}{(1-e_0^2)^2(1-5\cos^2 i_0)^2}\left[e_0^2\left(\frac{449}{8}-\frac{365}{12}\qquad C_1'\right)+e_0^4\left(\frac{229}{32}-\frac{649}{96}C_0'+\frac{915}{32}C_1'\right)-\frac{19}{24}e_0^6C_0'\right]\sin(2\xi-2\eta)$$

$$+\frac{k_2 h_0^4\cos^4 i_0}{(1-e_0^2)^2(1-5\cos^2 i_0)^2}\left[e_0^2\left(-\frac{2331}{16}+\frac{1275}{16}C_0'-\frac{405}{16}C_1'\right)-e_0^4\left(\frac{513}{32}-\frac{471}{32}C_0'+\frac{2295}{32}C_1'\right)+\frac{21}{8}e_0^6C_0'\right]\sin(2\xi-2\eta)$$

$$+\frac{k_2 h_0^4\cos^6 i_0}{(1-e_0^2)^2(1-5\cos^2 i_0)^2}\left[e_0^2\left(\frac{765}{8}-\frac{105}{2}C_0'+\frac{135}{8}C_1'\right)+e_0^4\left(\frac{315}{32}-\frac{285}{32}C_0'+\frac{1485}{32}C_1'\right)-\frac{15}{8}e_0^6C_0'\right]\sin(2\xi-2\eta)$$

$$+\frac{k_3^2(1-\cos^2 i_0)}{k_2^3(1-e_0^2)^2(1-5\cos^2 i_0)}\left(-\frac{29}{48}e_0^2-\frac{17}{48}e_0^4+\frac{23}{24}e_0^6\right)\sin(2\xi-2\eta)$$

$$+\frac{k_4 h_0^4(1-11\cos^2 i_0+31\cos^4 i_0-21\cos^6 i_0)}{k_2(1-e_0^2)^2(1-5\cos^2 i_0)^2}\left[e_0^2\left(\frac{25}{16}-\frac{65}{48}C_0'+\frac{15}{16}C_1'\right)-e_0^4\left(\frac{25}{16}-\frac{85}{48}C_0'+\frac{15}{16}C_1'\right)-\frac{5}{12}e_0^6C_0'\right]\sin(2\xi-2\eta)$$

$$+\frac{k_2 h_0^4 (1-2\cos^2 i_0+\cos^4 i_0)}{(1-e_0^2)^2(1-5\cos^2 i_0)}\left[-\frac{223}{256}e_0^4-\frac{169}{512}e_0^6+\frac{e_0^6}{1536}\left(25+\frac{7}{4}e_0^2\right)G-\frac{e_0^6}{768}\left(5+\frac{21}{2}e_0^2\right)H+\frac{11}{8192}e_0^8 I\right]\sin(4\xi-4\eta)$$

$$+\frac{k_2 h_0^4}{(1-e_0^2)^2(1-5\cos^2 i_0)^2}\left[\frac{209}{128}e_0^4+\frac{273}{1024}e_0^6+\frac{5}{3072}e_0^8-\left(\frac{25}{3072}e_0^6+\frac{43}{24576}e_0^8\right)G+\frac{5}{1536}e_0^8 H\right]\sin(4\xi-4\eta)$$

$$+\frac{k_2 h_0^4\cos^2 i_0}{(1-e_0^2)^2(1-5\cos^2 i_0)^2}\left[-\frac{1417}{128}e_0^4-\frac{2997}{1024}e_0^6-\frac{241}{3072}e_0^8+\left(\frac{25}{384}e_0^6+\frac{671}{24576}e_0^8\right)G-\frac{5}{192}e_0^8 H\right]\sin(4\xi-4\eta)$$

$$+\frac{k_2 h_0^4\cos^4 i_0}{(1-e_0^2)^2(1-5\cos^2 i_0)^2}\left[\frac{2207}{128}e_0^4+\frac{5175}{1024}e_0^6+\frac{467}{3072}e_0^8-\left(\frac{325}{3072}e_0^6+\frac{1213}{24576}e_0^8\right)G+\frac{65}{1536}e_0^8 H\right]\sin(4\xi-4\eta)$$

$$+\frac{k_2 h_0^4\cos^6 i_0}{(1-e_0^2)^2(1-5\cos^2 i_0)^2}\left[-\frac{999}{128}e_0^4-\frac{2451}{1024}e_0^6-\frac{77}{1024}e_0^8+\left(\frac{25}{512}e_0^6+\frac{195}{8192}e_0^8\right)G-\frac{5}{256}e_0^8 H\right]\sin(4\xi-4\eta)$$

$$+\frac{k_2 h_0^4}{(1-e_0^2)^2(1-5\cos^2 i_0)^3}\left(-\frac{265}{384}e_0^4-\frac{553}{6144}e_0^6-\frac{67}{6144}e_0^8\right)\sin(4\xi-4\eta)$$

$$+\frac{k_2 h_0^4\cos^2 i_0}{(1-e_0^2)^2(1-5\cos^2 i_0)^3}\left(\frac{5665}{768}e_0^4+\frac{701}{384}e_0^6+\frac{223}{768}e_0^8\right)\sin(4\xi-4\eta)$$

$$+\frac{k_2 h_0^4\cos^4 i_0}{(1-e_0^2)^2(1-5\cos^2 i_0)^3}\left(-\frac{18695}{768}e_0^4-\frac{30263}{3072}e_0^6-\frac{6737}{3072}e_0^8\right)\sin(4\xi-4\eta)$$

$$+\frac{k_2 h_0^4\cos^6 i_0}{(1-e_0^2)^2(1-5\cos^2 i_0)^3}\left(\frac{7505}{256}e_0^4+\frac{1867}{128}e_0^6+\frac{911}{256}e_0^8\right)\sin(4\xi-4\eta)$$

$$+\frac{k_2 h_0^4\cos^8 i_0}{(1-e_0^2)^2(1-5\cos^2 i_0)^3}\left(-\frac{2985}{256}e_0^4-\frac{13251}{2048}e_0^6-\frac{3369}{2048}e_0^8\right)\sin(4\xi-4\eta)$$

$$+\frac{k_4 h_0^4(1-9\cos^2 i_0+15\cos^4 i_0-7\cos^6 i_0)}{k_2(1-e_0^2)^2(1-5\cos^2 i_0)^2}\left[\frac{35}{64}e_0^4-\frac{705}{512}e_0^6-\frac{25}{1536}e_0^8+\frac{5}{3072}e_0^6\left(5+\frac{43}{4}e_0^2\right)G-\frac{5}{1536}e_0^8 H\right]\sin(4\xi-4\eta)$$

$$+\frac{k_4 h_0^4}{k_2(1-e_0^2)^2(1-5\cos^2 i_0)^3}\left(-\frac{875}{768}e_0^4+\frac{1415}{1536}e_0^6+\frac{335}{1536}e_0^8\right)\sin(4\xi-4\eta)$$

$$+\frac{k_4 h_0^4\cos^2 i_0}{k_2(1-e_0^2)^2(1-5\cos^2 i_0)^3}\left(\frac{2345}{128}e_0^4-\frac{875}{64}e_0^6-\frac{595}{128}e_0^8\right)\sin(4\xi-4\eta)$$

$$+\frac{k_4 h_0^4\cos^4 i_0}{k_2(1-e_0^2)^2(1-5\cos^2 i_0)^3}\left(-\frac{1435}{16}e_0^4+\frac{15935}{256}e_0^6+\frac{7025}{256}e_0^8\right)\sin(4\xi-4\eta)$$

$$+\frac{k_4 h_0^4\cos^6 i_0}{k_2(1-e_0^2)^2(1-5\cos^2 i_0)^3}\left(\frac{49525}{384}e_0^4-\frac{16735}{192}e_0^6-\frac{16055}{384}e_0^8\right)\sin(4\xi-4\eta)$$

$$+\frac{k_4 h_0^4\cos^8 i_0}{k_2(1-e_0^2)^2(1-5\cos^2 i_0)^3}\left(-\frac{14455}{256}e_0^4+\frac{19285}{512}e_0^6+\frac{9625}{512}e_0^8\right)\sin(4\xi-4\eta)$$

$$+\frac{k_4^2 h_0^4(1-16\cos^2 i_0+78\cos^4 i_0-112\cos^6 i_0+49\cos^8 i_0)}{k_2^3(1-e_0^2)^2(1-5\cos^2 i_0)^3}\left[\frac{125}{96}e_0^4-\frac{325}{1536}e_0^6-\frac{1675}{1536}e_0^8\right]\sin(4\xi-4\eta)$$

$$+\frac{k_3 h_0^2 \sin i_0}{k_2 (1-e_0^2)^2 (1-5\cos^2 i_0)}\left[e_0\left(\frac{9}{4}-\frac{13}{12}C_0'+\frac{3}{4}C_1'\right)+e_0^3\left(-\frac{9}{16}+\frac{17}{12}C_0'-\frac{3}{4}C_1'\right)+e_0^5\left(-\frac{289}{96}-\frac{1}{3}C_0'\right)+\frac{29}{32}e_0^7+\frac{11}{768}e_0^7 G\right]\cos(\xi-\eta)$$

$$+\frac{k_3 h_0^2 \sin i_0 \cos^2 i_0}{k_2 (1-e_0^2)^2 (1-5\cos^2 i_0)}\left[e_0\left(-\frac{19}{4}+\frac{13}{4}C_0'-\frac{9}{4}C_1'\right)+e_0^3\left(\frac{49}{16}-\frac{17}{4}C_0'+\frac{9}{4}C_1'\right)+e_0^5\left(\frac{289}{96}+C_0'\right)-\frac{29}{32}e_0^7-\frac{11}{768}e_0^7 G\right]\cos(\xi-\eta)$$

$$+\frac{k_3 h_0^2 \sin i_0}{k_2 (1-e_0^2)^2 (1-5\cos^2 i_0)^2}\left(-\frac{5}{4}e_0-\frac{125}{96}e_0^3+\frac{181}{96}e_0^5+\frac{2}{3}e_0^7\right)\cos(\xi-\eta)$$

$$+\frac{k_3 h_0^2 \sin i_0 \cos^2 i_0}{k_2 (1-e_0^2)^2 (1-5\cos^2 i_0)^2}\left(\frac{35}{4}e_0+\frac{221}{24}e_0^3-\frac{211}{24}e_0^5-\frac{55}{6}e_0^7\right)\cos(\xi-\eta)$$

$$+\frac{k_3 h_0^2 \sin i_0 \cos^4 i_0}{k_2 (1-e_0^2)^2 (1-5\cos^2 i_0)^2}\left(-\frac{15}{2}e_0-\frac{253}{32}e_0^3+\frac{221}{32}e_0^5+\frac{17}{2}e_0^7\right)\cos(\xi-\eta)$$

$$+\frac{k_3 k_4 h_0^2 \sin i_0 (1-8\cos^2 i_0+7\cos^4 i_0)}{k_2^3 (1-e_0^2)^2 (1-5\cos^2 i_0)^2}\left(\frac{5}{4}e_0+\frac{85}{48}e_0^3+\frac{175}{48}e_0^5-\frac{20}{3}e_0^7\right)\cos(\xi-\eta)$$

$$+\frac{k_3 h_0^2 \sin i_0 (1-\cos^2 i_0)}{k_2 (1-e_0^2)^2 (1-5\cos^2 i_0)}\left[\frac{13}{24}e_0^3-\frac{27}{32}e_0^5+\frac{17}{288}e_0^7+\frac{e_0^5}{1152}\left(5+\frac{29}{2}e_0^2\right)G-\frac{e_0^7}{576}H\right]\cos(3\xi-3\eta)$$

$$+\frac{k_3 h_0^2 \sin i_0}{k_2 (1-e_0^2)^2 (1-5\cos^2 i_0)^2}\left(-\frac{265}{288}e_0^3+\frac{209}{288}e_0^5+\frac{7}{36}e_0^7\right)\cos(3\xi-3\eta)$$

$$+\frac{k_3 h_0^2 \sin i_0 \cos^2 i_0}{k_2 (1-e_0^2)^2 (1-5\cos^2 i_0)^2}\left(\frac{511}{72}e_0^3-\frac{323}{72}e_0^5-\frac{47}{18}e_0^7\right)\cos(3\xi-3\eta)$$

$$+\frac{k_3 h_0^2 \sin i_0 \cos^4 i_0}{k_2 (1-e_0^2)^2 (1-5\cos^2 i_0)^2}\left(-\frac{593}{96}e_0^3+\frac{361}{96}e_0^5+\frac{29}{12}e_0^7\right)\cos(3\xi-3\eta)$$

$$+\frac{k_3 k_4 h_0^2 \sin i_0 (1-8\cos^2 i_0+7\cos^4 i_0)}{k_2^3 (1-e_0^2)^2 (1-5\cos^2 i_0)^2}\left(\frac{245}{144}e_0^3+\frac{35}{144}e_0^5-\frac{35}{18}e_0^7\right)\cos(3\xi-3\eta)$$

$$+\frac{k_2 h_0^4 \sin^4 i_0 (2-9\cos^2 i_0)(1-15\cos^2 i_0)}{(1-e_0^2)^2 (1-5\cos^2 i_0)^3}\left[+\frac{1}{32}e_0^4-\frac{27}{512}e_0^6+\frac{13}{1536}e_0^8+\frac{5}{12288}e_0^8 G\right]\sin(4\xi-4\eta)$$

$$+\frac{k_2 h_0^4 \sin^4 i_0 (2-9\cos^2 i_0)(1-15\cos^2 i_0)}{(1-e_0^2)(1-5\cos^2 i_0)^4}\left[-\frac{5}{128}e_0^4-\frac{5}{512}e_0^6+\left(\frac{15}{64}e_0^4+\frac{63}{512}e_0^6\right)\cos^2 i_0\right]\sin(4\xi-4\eta)$$

$$+\frac{k_2 h_0^4 \sin^4 i_0 (2-9\cos^2 i_0)^2 (1-15\cos^2 i_0)^2}{(1-e_0^2)(1-5\cos^2 i_0)^5}\frac{e_0^6}{512}\sin(4\xi-4\eta)$$

$$+\frac{k_3 h_0^2 \sin^3 i_0 (2-9\cos^2 i_0)}{k_2 (1-e_0^2)^2 (1-5\cos^2 i_0)^2}\left[-\frac{1}{3}e_0^3+\frac{9}{16}e_0^5-\frac{13}{144}e_0^7-\frac{5}{1152}e_0^7 G\right]\left[3\cos(\xi-\eta)+\cos(3\xi-3\eta)\right]$$

$$+\frac{k_3 h_0^2 \sin^3 i_0 (2-9\cos^2 i_0)}{k_2 (1-e_0^2)(1-5\cos^2 i_0)^3}\left[+\frac{7}{16}e_0^3+\frac{3}{16}e_0^5-\left(\frac{45}{16}e_0^3+\frac{41}{16}e_0^5\right)\cos^2 i_0\right]\left[3\cos(\xi-\eta)+\cos(3\xi-3\eta)\right]$$

$$-\frac{1}{24}\frac{k_3 h_0^2 e_0^5 \sin^3 i_0 (2-9\cos^2 i_0)^2 (1-15\cos^2 i_0)}{k_2 (1-e_0^2)(1-5\cos^2 i_0)^4}\left[3\cos(\xi-\eta)+\cos(3\xi-3\eta)\right]$$

$$+\frac{k_4 h_0^4 \sin^4 i_0 (2 - 9\cos^2 i_0)(1 - 7\cos^2 i_0)}{k_2 (1 - e_0^2)^2 (1 - 5\cos^2 i_0)^3}\left[-\frac{5}{16}e_0^4 + \frac{135}{256}e_0^6 - \frac{65}{768}e_0^8 - \frac{25}{6144}e_0^8 G\right] \sin(4\xi - 4\eta)$$

$$+\frac{k_4 h_0^4 \sin^4 i_0 (2 - 9\cos^2 i_0)(1 - 7\cos^2 i_0)}{k_2 (1 - e_0^2)(1 - 5\cos^2 i_0)^4}\left[\frac{55}{128}e_0^4 + \frac{25}{128}e_0^6 - \left(\frac{375}{128}e_0^4 + \frac{345}{128}e_0^6\right)\cos^2 i_0\right] \sin(4\xi - 4\eta)$$

$$-\frac{5}{128}\frac{k_4 h_0^4 \sin^4 i_0 (2 - 9\cos^2 i_0)^2 (1 - 15\cos^2 i_0)(1 - 7\cos^2 i_0)}{k_2 (1 - e_0^2)(1 - 5\cos^2 i_0)^5} e_0^6 \sin(4\xi - 4\eta)$$

$$-\frac{5}{8}\frac{k_3 k_4 h_0^2 \sin^3 i_0 (2 - 9\cos^2 i_0)(1 - 7\cos^2 i_0)}{k_2^3 (1 - e_0^2)(1 - 5\cos^2 i_0)^3}(e_0^3 + 3e_0^5)\left[3\cos(\xi - \eta) + \cos(3\xi - 3\eta)\right]$$

$$+\frac{5}{12}\frac{k_3 k_4 h_0^2 \sin^3 i_0 (2 - 9\cos^2 i_0)^2 (1 - 7\cos^2 i_0) e_0^5}{k_2^3 (1 - e_0^2)(1 - 5\cos^2 i_0)^4}\left[3\cos(\xi - \eta) + \cos(3\xi - 3\eta)\right]$$

$$+\frac{1}{4}\frac{k_3^2 \sin^2 i_0 (2 - 9\cos^2 i_0)}{k_2^3 (1 - e_0^2)(1 - 5\cos^2 i_0)^2}(e_0^2 + 4e_0^4)\sin(2\xi - 2\eta)$$

$$-\frac{1}{4}\frac{k_3^2 \sin^2 i_0 (2 - 9\cos^2 i_0)^2}{k_2^3 (1 - e_0^2)(1 - 5\cos^2 i_0)^3} e_0^4 \sin(2\xi - 2\eta)$$

$$-\frac{25}{128}\frac{k_4^2 h_0^4 \sin^4 i_0 (2 - 9\cos^2 i_0)(1 - 7\cos^2 i_0)^2}{k_2^3 (1 - e_0^2)(1 - 5\cos^2 i_0)^4}(2e_0^4 + 5e_0^6)\sin(4\xi - 4\eta)$$

$$+\frac{25}{128}\frac{k_4^2 h_0^4 \sin^4 i_0 (2 - 9\cos^2 i_0)^2 (1 - 7\cos^2 i_0)^2}{k_2^3 (1 - e_0^2)(1 - 5\cos^2 i_0)^5} e_0^6 \sin(4\xi - 4\eta) .$$

A SATELLITE ORBIT COMPUTATION PROGRAM FOR IZSAK'S SECOND-ORDER SOLUTION OF VINTI'S DYNAMICAL PROBLEM

RAYMOND V. BORCHERS

Goddard Space Flight Center

This report extends the results of Vinti and Izsak and presents a computational procedure designed specifically for Izsak's second-order solution of Vinti's dynamical problem. With this procedure, the coordinates and velocity of an unretarded satellite can be obtained from a knowledge of its initial conditions r and v.

In this procedure, the derivation is given for the complete set of six canonical constants from initial conditions. Three of these have been determined by Vinti and the remaining three by the author. All six of them are assumed known in Izsak's solution.

This report includes an adaptation of a Newton–Raphson iteration scheme specifically designed to solve a certain system of nonlinear equations introduced by Vinti for the purpose of numerically factoring a certain quartic equation. The solution by this method can be used instead of certain infinite series to obtain Izsak's elements a and e. An example is included to illustrate how these elements may be obtained by the Newton–Raphson method.

Appendix B gives the derivation of exact expressions for the components of velocity in Vinti's accurate intermediary satellite orbit using Izsak's orbital elements. The derivation is one of the necessary steps in comparing such a method with others.

INTRODUCTION

This report provides a computational procedure for determining the orbit of an artificial satellite in the earth's gravitational field. The procedure is based on Izsak's second-order solution of Vinti's dynamical problem.[1] This computing procedure differs from many other methods in that the potential function is included in an analytic solution of the equations of motion. This is advantageous because the difficulties associated with the slow convergence or divergence of some series expansions used in orbit calculations are avoided; also the problem of small divisors is avoided. Another advantage of this procedure is that it does not involve several multiplications of Fourier series, a task common to certain satellite programs. Although Fourier series are well adapted to numerical computation, it is certainly desirable from the standpoint of machine storage and computing time to minimize the total number of such series. In many satellite theories, Fourier series are used from the very beginning to obtain successive approximations of different orders to the solution. The use of Vinti's potential minimizes the use of perturbation theory; Izsak[1] states that the oblateness perturbations which are not accounted for by Vinti's potential can be treated by a first-order method, that is, without multiplications of Fourier series.

As was pointed out by Izsak[2] it is advantageous for several practical purposes to have satellite orbits with very small eccentricities. Since the eccentricity never appears as a divisor, this procedure is valid for arbitrarily small values of e or $e=0$. However, we must avoid polar orbits and orbits which have inclinations of less than 2 degrees.

Vinti[3] found an axially symmetric solution of Laplace's equation in oblate spheroidal coordinates which may be used as the gravitational potential about an oblate planet. This potential, which leads to separability of the Hamilton–

Jacobi equation, is a remarkable approximation to the actual gravitational field of the earth in that it fits the zeroth and second zonal harmonics exactly and accounts for over half of the fourth zonal harmonic. Naturally, the oblateness perturbations are only a part of the factors which affect the satellite motion. Other perturbations not accounted for in this procedure are the effects of the odd harmonics, the residual fourth harmonic, the lunar–solar forces, and aerodynamic and electromagnetic drag.

MATHEMATICAL PROBLEM

In Hamiltonian form, the equations of motion of a dynamical system of n degrees of freedom assume the forms

$$\left.\begin{aligned} \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i}, \\ \frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i}, \end{aligned}\right\} \qquad i = 1, 2, \cdots, n \tag{1}$$

where $H(q_1, q_2, \cdots, q_n; p_1, p_2, \cdots, p_n; t)$ is the Hamiltonian function (in which time appears explicitly) of the system with n generalized coordinates $q_1, q_2, \cdots, q_n$ and the conjugate momenta $p_1, p_2, \cdots, p_n$.

Solving the Hamilton-Jacobi equation

$$\frac{\partial \hat{W}}{\partial t} + H\left(q_1, q_2, \cdots, q_n; \frac{\partial \hat{W}}{\partial q_1}, \frac{\partial \hat{W}}{\partial q_2}, \cdots, \frac{\partial \hat{W}}{\partial q_n}\right) = 0,$$

where $\hat{W}$ is Hamilton's characteristic function, is equivalent to solving the Hamiltonian equations of motion (equation 1). If it is possible to separate the variables in the Hamilton-Jacobi equation, then the solution can always be reduced to quadratures.

Vinti's dynamical system belongs to a class of systems which are scleronomic, conservative, and holonomic. Furthermore, it belongs to a class of dynamical systems which are said to be of Stäckel's type. The separability properties of the Hamilton–Jacobi equation of the form solved by Vinti follow from certain conditions determined by Stäckel. The separability of the variables occurs only in certain coordinate systems.

The oblate spheroidal coordinate system is related to the geocentric rectangular coordinate system by

$$x = \sqrt{(\rho^2 + c^2)}\sqrt{1 - \sigma^2}\cos\alpha,$$

$$y = \sqrt{\rho^2 + c^2}\sqrt{1 - \sigma^2}\sin\alpha,$$

$$z = \rho\sigma,$$

$$r = \sqrt{\rho^2 + c^2(1 - \sigma^2)},$$

where x, y, and z are rectangular coordinates; r is the geocentric distance of the satellite; ρ, σ, and α are the coordinates in the oblate spheroidal system; and c is a constant defined by Vinti's expression

$$c^2 = J_2 a_E^2, \tag{2}$$

where J_2 is the coefficient of the second-degree Legendre polynomial in the earth's force function F. The quantity F is expressed as

$$F = +\frac{\mu}{r}\left[1 - \sum_{n=1}^{\infty} J_n\left(\frac{a_E}{r}\right)^n P_n(\sin\delta)\right],$$

where δ is declination of the satellite, a_E is the earth's equatorial radius, and μ is the product GM where G is the gravitational constant and M is the earth's mass.

The potential which Vinti obtained in oblate spheroidal coordinates is

$$\hat{V} = -\frac{\mu\rho}{\rho^2 + c^2\sigma^2}.$$

Similarly, the Hamiltonian and Lagrangian are

$$H = \frac{1}{2}U^2 - \frac{\mu\rho}{\rho^2 + c^2\sigma^2},$$

$$L = \frac{1}{2}U^2 + \frac{\mu\rho}{\rho^2 + c^2\sigma^2},$$

where the speed of the satellite U is found from

$$U^2 = \frac{\rho^2 + c^2\sigma^2}{\rho^2 + c^2}\dot{\rho}^2 + \frac{\rho^2 + c^2\sigma^2}{1 - \sigma^2}\dot{\sigma}^2 + \left(\rho^2 + c^2\right)\left(1 - \sigma^2\right)\dot{\alpha}^2 .$$

The generalized momenta are defined by

$$\left.\begin{aligned} P_\rho &= \frac{\partial L}{\partial \dot{\rho}} , \\ P_\sigma &= \frac{\partial L}{\partial \dot{\sigma}} , \\ P_\alpha &= \frac{\partial L}{\partial \dot{\alpha}} , \end{aligned}\right\}$$

The Hamiltonian does not contain the time explicitly, so the Hamilton–Jacobi equation is

$$\frac{1}{2\left(\rho^2 + c^2\sigma^2\right)}\left[\left(\rho^2 + c^2\right)\left(\frac{\partial \hat{W}}{\partial \rho}\right)^2 + \left(1 - \sigma^2\right)\left(\frac{\partial \hat{W}}{\partial \sigma}\right)^2 + \left(\frac{1}{1 - \sigma^2} - \frac{c^2}{\rho^2 + c^2}\right)\left(\frac{\partial \hat{W}}{\partial \alpha}\right)^2\right] - \frac{\mu\rho}{\rho^2 + c^2\sigma^2} = \hat{h} ,$$

where, in the limit as $c^2 \to 0$ of Keplerian motion, $\hat{h}$ is the total energy in the orbit; $\hat{h}$ is always negative.

The implicit equations of motion for the determination of ρ, σ, and α are[1]

$$\left.\begin{aligned} \frac{\partial \hat{W}}{\partial \hat{h}} &= \int_{\rho_1}^{\rho} \frac{\rho^2\, d\rho}{\sqrt{P(\rho)}} + c^2 \int_0^{\sigma} \frac{\sigma^2\, d\sigma}{\sqrt{Q(\sigma)}} = t - \hat{t} , \\ \frac{\partial \hat{W}}{\partial \hat{c}} &= -\hat{c}\int_{\rho_1}^{\rho} \frac{d\rho}{\sqrt{P(\rho)}} + \hat{c}\int_0^{\sigma} \frac{d\sigma}{\sqrt{Q(\sigma)}} = \hat{\omega} , \\ \frac{\partial \hat{W}}{\partial \hat{G}} &= c^2 \hat{G}\int_{\rho_1}^{\rho} \frac{d\rho}{\left(\rho^2 + c^2\right)\sqrt{P(\rho)}} - \hat{G}\int_0^{\sigma} \frac{d\sigma}{\left(1 - \sigma^2\right)\sqrt{Q(\sigma)}} + \alpha = \hat{\Omega} , \end{aligned}\right\}$$

where

$$\left.\begin{aligned} P(\rho) &= 2\hat{h}\rho^4 + 2\mu\rho^3 - \left(\hat{c}^2 - 2c^2\hat{h}\right)\rho^2 + 2c^2\mu\rho - c^2\left(\hat{c}^2 - \hat{G}^2\right) , \\ Q(\sigma) &= -2c^2\hat{h}\sigma^4 - \left(\hat{c}^2 - 2c^2\hat{h}\right)\sigma^2 + \left(\hat{c}^2 - \hat{G}^2\right) . \end{aligned}\right\} \quad (3)$$

The symbols $\hat{h}$, $\hat{c}$, $\hat{G}$, $-\hat{t}$, $\hat{\omega}$, and $\hat{\Omega}$ are a canonical set of constants of integration. In the limit as $c^2 \to 0$ of Keplerian motion, the canonical constants have the following meanings:

- $\hat{h}$ total energy in the orbit, $\hat{h}$ is always negative;
- $\hat{c}$ total angular momentum;
- $\hat{G}$ z component of the angular momentum, $\hat{G}$ is positive or negative according as the motion is direct or retrograde;
- $-\hat{t}$ time of perigee passage;
- $\hat{\omega}$ argument of perigee; and
- $\hat{\Omega}$ right ascension of the ascending node.

Exact expressions for three canonical constants

α_1, α_2, and α_3, denoted by Izsak as $\hat{h}$, $\hat{c}$, and $\hat{G}$, respectively, are determined from initial values of the coordinates and their derivatives.[4] Numerical values of these α's are used to determine a certain set of orbital constants a_0, e_0, and i_0 (the initial values of the semimajor axis of the orbit, the eccentricity of the orbit, and the angle of inclination, respectively). These are used to find the ρ_1, ρ_2, A, and B (the perigee of the orbit, the apogee of the orbit, and coefficients in the quartic polynomial $F(\rho)$—see Appendix B—respectively) necessary to factor

$$F(\rho) = -2\alpha_1 (\rho - \rho_1)(\rho_2 - \rho)(\rho^2 + A\rho + B) , \quad (4)$$

where $\rho_1 = a(1-e)$ and $\rho_2 = a(1+e)$. Similarly, this same quartic designated by Izsak as $P(\rho)$, equation 3, can be factored into a form which is equivalent to that of $F(\rho)$, equation 4. That is,

$$P(\rho) = -2\hat{h}(\rho_2 - \rho)(\rho - \rho_1)\left[(\rho - a\kappa)^2 + a^2\lambda^2\right] ,$$

and we find that

$$\rho^2 + A\rho + B = (\rho - a\kappa)^2 + a^2\lambda^2 .$$

The values for $\rho_1+\rho_2$, $\rho_1\rho_2$, A, and B are determined initially by solving the following system of nonlinear equations:

$$\left.\begin{aligned}
&\rho_1 + \rho_2 - A = -\mu\alpha_1^{-1} = 2a_0 ,\\
&B + \rho_1\rho_2 - (\rho_1 + \rho_2)A = c^2 - \tfrac{1}{2}\alpha_2^2\alpha_1^{-1} = c^2 + a_0 p_0 ,\\
&(\rho_1 + \rho_2)B - \rho_1\rho_2 A = -\mu\alpha_1^{-1}c^2 = 2a_0 c^2 ,\\
&\rho_1\rho_2 B = -\tfrac{1}{2}c^2(\alpha_2^2 - \alpha_3^2)\alpha_1^{-1} = a_0 p_0 c^2 \sin^2 i_0 .
\end{aligned}\right\}$$

Vinti[4] has given a second-order solution of this system by a method of successive approximations. However, if higher order accuracy is desired, it is first necessary to obtain additional terms in the series solutions; this is a laborious task. A numerical method to obtain the solution is given in the next section.

NEWTON-RAPHSON ITERATION SCHEME

The solution of a set of nonlinear algebraic equations usually involves a great deal more work than that needed for linear systems. When n, the number of equations, is large, the solution of linear systems entails considerable computation time even on high-speed computers; the solution of nonlinear systems may often be almost prohibitive.

The Newton-Raphson method[5,6,7] can easily be applied when a solution is required for only a few equations.

To solve a system of nonlinear equations such as

$$(\rho_1 + \rho_2) - A = 2a_0 , \quad (5)$$

$$B + \rho_1\rho_2 - (\rho_1 + \rho_2)A = c^2 + a_0 p_0 , \quad (6)$$

$$(\rho_1 + \rho_2)B - \rho_1\rho_2 A = 2a_0 c^2 , \quad (7)$$

$$\rho_1\rho_2 B = a_0 p_0 c^2 \sin^2 i_0 , \quad (8)$$

with unknowns $(\rho_1+\rho_2)$, $\rho_1\rho_2$, A, and B by the Newton-Raphson iteration we begin with a trial vector

$$X^{(k+1)} = X^{(k)} - J^{-1}F(X^{(k)}) ,$$

$$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \rho_1 + \rho_2 \\ \rho_1\rho_2 \\ A \\ B \end{pmatrix} = \begin{pmatrix} 2a_0 \\ a_0 p_0 \\ 0 \\ 0 \end{pmatrix} . \quad (9)$$

Denoting equations 5, 6, 7, and 8 by f_1, f_2, f_3, and f_4, respectively, that is,

$$f_1(x_1, x_2, x_3, x_4) = (\rho_1 + \rho_2) - A - 2a_0 ,$$

$$f_2(x_1, x_2, x_3, x_4) = B + \rho_1 \rho_2 - (\rho_1 + \rho_2) A - c^2 - a_0 p_0 ,$$

$$f_3(x_1, x_2, x_3, x_4) = (\rho_1 + \rho_2) B - \rho_1 \rho_2 A - 2a_0 c^2 ,$$

$$f_4(x_1, x_2, x_3, x_4) = \rho_1 \rho_2 B - a_0 p_0 c^2 \sin^2 i_0 ,$$

we introduce the usual Jacobian matrix

$$J = \left(\frac{\partial (f_1, f_2, f_3, f_4)}{\partial (x_1, x_2, x_3, x_4)} \right)$$

$$= \begin{pmatrix} 1 & 0 & -1 & 0 \\ -A & 1 & -(\rho_1 + \rho_2) & 1 \\ B & -A & -\rho_1 \rho_2 & \rho_1 + \rho_2 \\ 0 & B & 0 & \rho_1 \rho_2 \end{pmatrix} . \qquad (10)$$

It will be noted that for the initial vector, the Jacobian determinant can be written

$$|J| = \begin{vmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -2a_0 & 1 \\ 0 & 0 & -a_0 p_0 & 2a_0 \\ 0 & 0 & 0 & a_0 p_0 \end{vmatrix} = -a_0^2 p_0^2 = -a_0^4 (1 - e_0^2)^2 .$$

The condition that $| J |$ is not close to zero will be satisfied provided e_0 is not close to unity.

Next, we determine the exact inverse of equation 10; only the final result is given here:

$$J^{-1} = \begin{pmatrix} Z_{11} & Z_{12} & Z_{13} & Z_{14} \\ Z_{21} & Z_{22} & Z_{23} & Z_{24} \\ Z_{31} & Z_{32} & Z_{33} & Z_{34} \\ Z_{41} & Z_{42} & Z_{43} & Z_{44} \end{pmatrix} ,$$

where

$$Z_{11} = 1 - \left[\frac{(\rho_1 \rho_2 - B)(B - A^2) - BA (A + \rho_1 + \rho_2)}{\Delta} \right] ,$$

$$Z_{12} = \left[\frac{\rho_1 \rho_2 A + (\rho_1 + \rho_2) B}{\Delta} \right] ,$$

$$Z_{13} = \frac{\rho_1 \rho_2 - B}{\Delta} ,$$

$$Z_{14} = -\frac{A + \rho_1 + \rho_2}{\Delta} ,$$

$$Z_{21} = A - \left\{ \frac{(\rho_1 \rho_2)(A+\rho_1+\rho_2)(B-A^2) - BA(A+\rho_1+\rho_2)^2 + BA\left[(\rho_1\rho_2 - B) + A(A+\rho_1+\rho_2)\right]}{\Delta} \right\} ,$$

$$Z_{22} = 1 - \left\{ \frac{(\rho_1 \rho_2 - B)\left[-A(A+\rho_1+\rho_2) + B\right] - B(A+\rho_1+\rho_2)^2}{\Delta} \right\} ,$$

$$Z_{23} = -(\rho_1 \rho_2) Z_{14} = \frac{(\rho_1 \rho_2)(A+\rho_1+\rho_2)}{\Delta} ,$$

$$Z_{24} = \frac{-(A+\rho_1+\rho_2)^2 + (\rho_1\rho_2 - B) + A(A+\rho_1+\rho_2)}{\Delta} ,$$

$$Z_{31} = (Z_{11} - 1) = -\left[\frac{(\rho_1\rho_2 - B)(B-A^2) - BA(A+\rho_1+\rho_2)}{\Delta} \right] ,$$

$$Z_{32} = Z_{12} = \frac{(\rho_1 \rho_2)A + (\rho_1+\rho_2)B}{\Delta} ,$$

$$Z_{33} = Z_{13} = \frac{(\rho_1\rho_2 - B)}{\Delta} ,$$

$$Z_{34} = Z_{14} = \frac{-(A+\rho_1+\rho_2)}{\Delta} ,$$

$$Z_{41} = BZ_{32} = \frac{B^2(\rho_1+\rho_2) + BA(\rho_1\rho_2)}{\Delta} ,$$

$$Z_{42} = BZ_{13} = \frac{B(\rho_1\rho_2 - B)}{\Delta} ,$$

$$Z_{43} = BZ_{34} = \frac{-B(A+\rho_1+\rho_2)}{\Delta} ,$$

$$Z_{44} = -Z_{33} + AZ_{34} = -\left[\frac{(\rho_1\rho_2 - B) + A(A+\rho_1+\rho_2)}{\Delta} \right] ,$$

where

$$\Delta = -\left[(\rho_1\rho_2 - B)^2 + A(\rho_1\rho_2 - B)(A+\rho_1+\rho_2) + B(A+\rho_1+\rho_2)^2 \right] .$$

The Newton-Raphson iteration can now be written

$$\begin{pmatrix} X_1^{(k+1)} \\ X_2^{(k+1)} \\ X_3^{(k+1)} \\ X_4^{(k+1)} \end{pmatrix} = \begin{pmatrix} X_1^{(k)} \\ X_2^{(k)} \\ X_3^{(k)} \\ X_4^{(k)} \end{pmatrix} - \begin{pmatrix} Z_{11} & Z_{12} & Z_{13} & Z_{14} \\ Z_{21} & Z_{22} & Z_{23} & Z_{24} \\ Z_{31} & Z_{32} & Z_{33} & Z_{34} \\ Z_{41} & Z_{42} & Z_{43} & Z_{44} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} .$$

A solution will have been obtained when

$$\operatorname*{Max}_i \left| \left(X_i^{(k+1)} - X_i^{(k)} \right) \right| \leq \epsilon ,$$

where ϵ is any tolerance sufficiently small to obtain the degree of accuracy desired.

The first-order solution through k_0 is obtained in one iteration by beginning with the zero-order solution of equation 9. It is expected that four iterations will be sufficient to obtain the solution through $0(k_0^3)$.

Table 1 gives the solution to the system as computed on the IBM 7090 using a single precision floating point Fortran program. The solution of equations 5 through 8 was obtained accurately to seven significant digits in four iterations with $\epsilon = 10^{-7}$.

The initial values of the unknowns, together with other necessary constants were computed from orbital data on Explorer XI (1961 ν).

We immediately obtain the values of the elements a and e from

$$a = \frac{\rho_1 + \rho_2}{2} ,$$

$$e = \sqrt{1 - \frac{4\rho_1 \rho_2}{(\rho_1 + \rho_2)^2}}$$

DETERMINATION OF CANONICAL CONSTANTS FROM INITIAL CONDITIONS

If the initial conditions (denoted by zero subscripts) t_0, x_0, y_0, z_0, $\dot{x}_0$, $\dot{y}_0$, and $\dot{z}_0$ are known, we can determine a complete set of canonical constants a, e, S, $-\hat{t}$, $\hat{\Omega}$, and ω essential to Izsak's second-order solution. The canonical constants have the following meanings:

a	semimajor axis of the orbit,
e	eccentricity of the orbit,
S	sine of the inclination of the orbit,

TABLE 1.—*Numerical Results Obtained with the Initial Conditions*

$\rho_1 + \rho_2 = 2a_0 = 15.0588664,$

$\rho_1 \rho_2 = a_0 p_0 = 56.2660106,$

$A = 0.0,$

$B = 0.0,$

$c^2 = J_2 a_E^2 = 0.044029034,$

$\sin i_0 = 0.474484778,$

$\epsilon = 1.0 \times 10^{-7}.$

Iteration Number	$\rho_1 + \rho_2$	$\rho_1 \rho_2$	A	B
1	15.0497347	56.1626263	-0.00913084351	+0.00991251186
2	15.0497213	56.1624885	-0.00914439194	+0.00993078329
3	15.0497213	56.1624877	-0.00914439194	+0.00993078336
4	15.0497213	56.1624877	-0.00914439209	+0.00993078336

$-\hat{t}$ in the limit as $c^2 \to 0$ of Keplerian motion $-\hat{t}$ is the time of perigee passage,

$\hat{\Omega}$ in the limit as $c^2 \to 0$ of Keplerian motion $\hat{\Omega}$ is the right ascension of the ascending node,

ω a constant of integration.

We first give the method of determining a, e, and S. The following expressions are computed:

$$r_0^2 = x_0^2 + y_0^2 + z_0^2 ,$$

$$v_0^2 = \dot{x}_0^2 + \dot{y}_0^2 + \dot{z}_0^2 ,$$

$$\rho_0^2 = \frac{r_0^2 - c^2}{2}\left[1 + \sqrt{1 + \left(\frac{2cz_0}{r_0^2 - c^2}\right)^2}\right],$$

$$\sigma_0^2 = \frac{2z_0^2}{r_0^2 - c^2}\left[1 + \sqrt{1 + \left(\frac{2cz_0}{r_0^2 - c^2}\right)^2}\right]^{-1} ,$$

$$r_0 \dot{r}_0 = x_0 \dot{x}_0 + y_0 \dot{y}_0 + z_0 \dot{z}_0 ,$$

$$\dot{\rho}_0 = \frac{\rho_0 r_0 \dot{r}_0 + c^2 \sigma_0 \dot{z}_0}{\rho_0^2 + c^2 \sigma_0^2} ,$$

$$\dot{\sigma}_0 = \frac{-\sigma_0 r_0 \dot{r}_0 + \rho_0 \dot{z}_0}{\rho_0^2 + c^2 \sigma_0^2} , \qquad (11)$$

where σ_0 takes the sign of z_0. Next, the following expressions are computed:

$$a_1 = \frac{1}{2} v_0^2 - \frac{\mu \rho_0}{\left(\rho_0^2 + c^2 \sigma_0^2\right)} , \quad a_1 < 0 ,$$

$$a_3 = x_0 \dot{y}_0 - y_0 \dot{x}_0 ,$$

$$a_2^2 = \frac{a_3^2 + \left(-\sigma_0 r_0 \dot{r}_0 + \rho_0 \dot{z}_0\right)^2}{1 - \sigma_0^2} - 2a_1 c^2 \sigma_0^2 ,$$

$$y^2 = \frac{a_3^2}{a_2^2} = \cos^2 i_0 ,$$

$$\sqrt{1 - y^2} = \sin i_0 ,$$

$$P_0 = \frac{a_2^2}{\mu} ,$$

$$K_0 = \frac{c^2}{P_0^2} ,$$

$$\rho_1 + \rho_2 = 2P_0 x^{-2}\left[1 - K_0 x^2 y^2 - K_0^2 x^2 y^2 \left(2x^2 - 3x^2 y^2 - 4 + 8y^2\right) + \cdots\right] ,$$

$$\rho_1 \rho_2 = P_0^2 x^{-2}\left[1 + K_0 y^2 \left(x^2 - 4\right) - K_0^2 y^2 \left(12x^2 - x^4 - 20x^2 y^2 - 16 + 32y^2 + x^4 y^2\right) + \cdots\right] ,$$

$$a = \frac{1}{2}(\rho_1 + \rho_2) ,$$

$$1 - e^2 = \frac{4\rho_1 \rho_2}{(\rho_1 + \rho_2)^2} ,$$

$$e = \frac{(\rho_2 - \rho_1)}{\rho_1 + \rho_2} = \sqrt{1 - (1 - e^2)} ,$$

$$\eta_0 = (\sin i_0)\left[1 - \frac{1}{2} K_0 x^2 y^2 + \frac{1}{8} K_0^2 x^4 y^2 (7y^2 - 4) + \cdots\right] , \tag{12}$$

where $\eta_0 \approx \sin I = S$. Alternately η_0 may be computed from

$$\eta_0^{-2} = \frac{\alpha_2^2 - 2\alpha_1 c^2}{2(\alpha_2^2 - \alpha_3^2)} \left[1 + \sqrt{1 + \frac{8\alpha_1 c^2 (\alpha_2^2 - \alpha_3^2)}{(\alpha_2^2 - 2\alpha_1 c^2)^2}}\right] .$$

We now have determined Izsak's elements a, e, and S from initial conditions; these elements are accurate through $0(k_0^2)$ and are used as input to the orbit computation procedure.

We are now ready to determine the remaining canonical constants $-\hat{\imath}$, $\hat{\Omega}$, and ω. We set $\phi = E = 0$ whenever $e = 0$. If $e \neq 0$, we determine E from

$$E = \tan^{-1}\left[\frac{\rho_0 r_0 \dot{r}_0 + c^2 \sigma_0 \dot{z}_0}{\sqrt{-2\hat{h}}\ \sqrt{(\rho - a\kappa)^2 + a^2\lambda^2}\ (a - \rho_0)}\right] ,$$

since

$$\cos E = \frac{a - \rho_0}{ae} ,$$

$$\sin E = \frac{\rho_0 r_0 \dot{r}_0 + c^2 \sigma_0 \dot{z}_0}{ae\sqrt{-2\hat{h}}\ \sqrt{(\rho - a\kappa)^2 + a^2\lambda^2}} ,$$

where ρ_0^2 is given by equation 11, c is given by equation 2, and

$$a^2\lambda^2 = a^2\left[\nu^2 S^2 - \frac{\nu^4}{(1 - e^2)^2}(1 - S^2)(1 - 5S^2) + \cdots\right] , \tag{13}$$

$$\kappa = \frac{\nu^2(1 - S^2)}{1 - e^2} + \frac{\nu^4(1 - S^2)}{(1 - e^2)^3}(1 - 4S^2 - e^2) + \cdots , \tag{14}$$

$$-2\hat{h} = \frac{\mu}{a(1 + \kappa)} . \tag{15}$$

The angle ϕ is completely determined by

$$\cos\phi = \frac{\cos E - e_*}{1 - e_* \cos E} , \tag{16}$$

$$\sin \phi = \frac{\sqrt{1 - e_*^2}\, \sin E}{1 - e_* \cos E}, \tag{17}$$

where e_* is given by

$$e_* = e\left\{1 + \frac{\nu^2}{1 - e^2}(1 - 2S^2) + \frac{\nu^4}{(1-e^2)^3}\left[(3 - 16S^2 + 14S^4) - 2(1 - S^2)^2 e^2\right] + \cdots\right\}. \tag{18}$$

We next determine the angle ψ from

$$\sin \psi = \frac{\sigma_0}{S}, \quad S \neq 0, \tag{19}$$

$$\cos \psi = \frac{-\sigma_0 r_0 \dot{r}_0 + \rho_0 \dot{z}_0}{a\sqrt{-2\hat{h}}\,\sqrt{1 - e^2}\,\sqrt{\frac{\kappa^2 + \lambda^2}{\nu^2}}\,\sqrt{1 - \ell^2 \sin^2 \psi}}, \tag{20}$$

where $(1-e^2)$ is given by equation 12 and

$$\ell^2 = \frac{\nu^2 S^2}{1 - e^2} - \frac{4\nu^4 S^2}{(1-e^2)^3}(1 - S^2) + \cdots, \tag{21}$$

$$\frac{\kappa^2 + \lambda^2}{\nu^2} = S^2 + \frac{4\nu^2 S^2}{(1-e^2)^2}(1 - S^2) + \frac{4\nu^4 S^2}{(1-e^2)^4}(1 - S^2)\left[(1 - 3S^2) - (1 + S^2)e^2\right] + \cdots. \tag{22}$$

Next we compute a "mean anomaly" $\hat{M}$ from

$$\hat{M} = E - K_1 e \sin E - K_2 \phi - K_3 \sin \phi + K_4 \sin 2\phi + K_5 \psi - K_6 \sin 2\psi + K_7 \sin 4\psi,$$

where

$$K_1 = 1 - \frac{\nu^2(1 - S^2)}{1 - e^2} + \frac{\nu^4(1 - S^2)}{(1 - e^2)^3} S^2 (3 + e^2), \tag{23}$$

$$K_2 = \frac{\nu^2 \sqrt{1 - e^2}}{2(1 - e)} S^2 - \frac{\nu^4 \sqrt{1 - e^2}}{16(1 - e^2)^3}\left[(24 - 96S^2 + 78S^4) - (8 - 11S^2) S^2 e^2\right], \tag{24}$$

$$K_3 = \frac{\nu^4 \sqrt{1 - e^2}}{4(1 - e^2)^3}(4 - 5S^2) S^2 e, \tag{25}$$

$$K_4 = \frac{3\nu^4 \sqrt{1 - e^2}\, S^4 e^2}{32(1 - e^2)^3}, \tag{26}$$

$$K_5 = \left\{\frac{\nu^2 \sqrt{1 - e^2}}{2(1 - e^2)} - \frac{\nu^4 \sqrt{1 - e^2}}{16(1 - e^2)^3}\left[(24 - 27S^2) - (8 - 11S^2) e^2\right]\right\} S^2. \tag{27}$$

$$K_6 = \left\{ \frac{\nu^2 \sqrt{1-e^2}}{4(1-e^2)} - \frac{\nu^4 \sqrt{1-e^2}}{8(1-e^2)^3} \left[(6-7S^2) - (2-3S^2)e^2 \right] \right\} S^2 , \tag{28}$$

$$K_7 = \frac{\nu^4 \sqrt{1-e^2}}{64(1-e^2)^2} S^4 . \tag{29}$$

We can now compute $-\hat{t}$ as follows:

$$-\hat{t} = \frac{\hat{M}}{\hat{n}} - t ,$$

where

$$\hat{n} = \sqrt{\frac{\mu}{a^3}} \left\{ 1 - \frac{3\nu^2(1-S^2)}{2(1-e^2)} + \frac{3\nu^4(1-S^2)}{8(1-e^2)^3} \left[(1+11S^2) - (1-5S^2)e^2 \right] - \cdots \right\} . \tag{30}$$

The right ascension of the satellite α is determined from

$$\cos\alpha = \frac{x_0}{\sqrt{\rho_0^2 + c^2}\ \sqrt{1-\sigma_0^2}} ,$$

$$\sin\alpha = \frac{y_0}{\sqrt{\rho_0^2 + c^2}\ \sqrt{1-\sigma_0^2}} .$$

When the right ascension α is known, the right ascension of the ascending node $\hat{\Omega}$ is computed from

$$\hat{\Omega} = \alpha - \tan^{-1}\left(\sqrt{1-S^2} \tan\psi \right) + R_1\psi - R_2 \sin 2\psi + R_3\phi$$

$$+ R_4 \sin\phi + R_5 \sin 2\phi - R_6 \sin 3\phi - R_7 \sin 4\phi ,$$

where

$$R_1 = \frac{\nu^2 \sqrt{1-S^2}}{2(1-e^2)} - \frac{\nu^4 \sqrt{1-S^2}}{16(1-e^2)^3} \left[(30-35S^2) + (2+3S^2)e^2 \right] , \tag{31}$$

$$R_2 = \frac{3\nu^4 \sqrt{1-S^2}}{32(1-e^2)^2} S^2 , \tag{32}$$

$$R_3 = \frac{\nu^2 \sqrt{1-S^2}}{2(1-e^2)^2}(2+e^2) + \frac{\nu^4 \sqrt{1-S^2}}{16(1-e^2)^4} \left[(24-56S^2) - (4+64S^2)e^2 - (2+3S^2)e^4 \right] , \tag{33}$$

$$R_4 = \left\{ \frac{2\nu^2 \sqrt{1-S^2}}{(1-e^2)^2} + \frac{\nu^4 \sqrt{1-S^2}}{4(1-e^2)^4} \left[(4-28S^2) - (6+7S^2)e^2 \right] \right\} e , \tag{34}$$

$$R_5 = \left\{ \frac{\nu^2 \sqrt{1 - S^2}}{4(1-e^2)^2} - \frac{\nu^4 \sqrt{1 - S^2}}{8(1-e^2)^4} \left[11 + (1 + S^2) e^2 \right] \right\} e^2 , \tag{35}$$

$$R_6 = \frac{\nu^4 \sqrt{1 - S^2}}{4(1-e^2)^4} (2 - S^2) e^3 , \tag{36}$$

$$R_7 = \frac{\nu^4 \sqrt{1 - S^2}}{64(1-e^2)^4} (2 + S^2) e^4 . \tag{37}$$

We next compute W and V, which are analogous to the argument of latitude in Keplerian motion and the true anomaly in Keplerian motion, respectively, from

$$W = \psi - M_1 \sin 2\psi + 3M_2 \sin 4\psi ,$$

$$V = \phi + L_1 \sin 2\phi + L_2 \sin 4\phi , \tag{38}$$

where

$$M_1 = \frac{\ell^2}{8} \left(1 + \frac{\ell^2}{2} \right) ,$$

$$M_2 = \frac{\ell^4}{256} ,$$

$$L_1 = \frac{-k^2}{8} \left(1 + \frac{k^2}{2} \right) ,$$

$$L_2 = \frac{3k^4}{256} .$$

The mean argument of perigee $\bar{\omega}$ is given by

$$\bar{\omega} = W - V .$$

The constant of integration ω is given by

$$\omega = W - (1 + \epsilon) V ,$$

where

$$\epsilon = \frac{\nu^2}{4(1-e^2)^2} (12 - 15S^2) + \frac{\nu^4}{64(1-e^2)^4} \left[(288 - 1296S^2 + 1035S^4) - (144 + 288S^2 - 510S^4) e^2 \right] + \cdots . \tag{39}$$

ORBIT COMPUTATION PROCEDURE

With the exception of the velocity formulation, the computational procedure developed here makes use of the unmodified expressions of Izsak.[1]

Input

The 11 inputs are: a, e, S, $-\hat{t}$, $\hat{\Omega}$, ω, J_2, a_E, μ, Δt, and T_f. The six inputs a, e, S, $-\hat{t}$, $\hat{\Omega}$, and ω are constants of integration (see definitions on p. 192). The other inputs have the following meanings:

- J_2 the coefficient of the second-degree Legendre polynomial in the earth's gravitational potential,
- a_E the earth's equatorial radius,
- μ the product GM where G is the gravitational constant and M is the earth's mass,
- Δt time interval of integration,
- T_f final time.

The following values of μ, J_2, and a_E determined by W. M. Kaula[8] were used in the computations:

$\mu = 3.986032 \times 10^2$ megameters^3 ksec^{-2},

$J_2 = 1.0823 \times 10^{-3}$,

$a_E = 6.378165$ megameters.

Equations and Fundamental Constants

From Vinti's expression (equation 2) and the input constants determined by Kaula, we have $c = 0.20983097$ megameters. In addition to $\alpha^2\lambda^2$ (equation 13), κ (equation 14), $-2h$ (equation 15), $\cos\phi$ (equation 16), $\sin\phi$ (equation 17), e_* (equation 18), l^2 (equation 21), $(\kappa^2+\lambda^2)/\nu^2$ (equation 22), $\hat{n}$ (equation 30), V (equation 38), and ϵ (equation 39), the following equations are used in the computation:

$$\nu^2 = \frac{c^2}{a^2},$$

$$k^2 = \frac{\nu^2 e^2}{(1-e^2)^2} S^2 - \frac{\nu^4 e^2}{(1-e^2)^4}\left[(1 - 10S^2 + 11S^4) + S^4 e^2\right] + \cdots,$$

and the generalized Kepler equation

$$E - K_1 e \sin E = \hat{n}(t - \hat{t}) + K_2\phi + K_3 \sin\phi - K_4 \sin 2\phi - K_5 \psi + K_6 \sin 2\psi - K_7 \sin 4\psi, \tag{40}$$

where the K_i are given by equations 23 through 29.

The right ascension α is computed from

$$\alpha = \hat{\Omega} + \tan^{-1}\left(\sqrt{1 - S^2}\tan\psi\right) - R_1\psi + R_2 \sin 2\psi - R_3\phi - R_4 \sin\phi$$

$$- R_5 \sin 2\phi + R_6 \sin 3\phi + R_7 \sin 4\phi,$$

where the R_i are given by equations 31 through 37.

The argument of latitude ψ is computed from the following equations:

$$W = (1+\epsilon)V + \omega,$$

$$\psi = W + M_1 \sin 2W + M_2 \sin 4W + \cdots.$$

The mean argument of perigee $\bar{\omega}$ is computed from

$$\bar{\omega} = \epsilon V + \omega.$$

The anomalistic mean motion n_ϕ is computed from

$$n_\phi = \hat{n}\left[1 - \frac{3\nu^4\sqrt{1-e^2}}{16(1-e^2)^3}(8 - 32S^2 + 25S^4) + \cdots\right] .$$

The motion of the node η is computed from

$$\eta = -\frac{3\nu^2\sqrt{1-S^2}}{2(1-e^2)^2} + \frac{3\nu^4\sqrt{1-S^2}}{16(1-e^2)^4}\left[(18 - 13S^2) + 24S^2e^2\right] - \cdots .$$

The oblate spheroidal coordinate σ is computed from

$$\sigma = s\sin\psi .$$

The z component of the angular momentum $\hat{G}$ is computed from

$$\hat{G}^2 = -2\hat{h}a^2(1-S^2)\left[(1-e^2)\left(\frac{\kappa^2+\lambda^2}{\nu^2 S^2}\right) - \nu^2\right] .$$

The oblate-spheroidal coordinate ρ is computed from

$$\rho = a(1 - e\cos E) .$$

Initially for $t=t_0$ we start with values ϕ and ψ determined from equations 16, 17, 19, and 20 to solve the generalized Kepler equation given by equation 40 using a Newton–Raphson iteration scheme. We test $|E(\phi_{i+1}, \psi_{i+1}) - E(\phi_i, \psi_i)| < \epsilon$, where $\epsilon > 0$ was chosen to be 10^{-7}. In general, only two or three iterations are required before sufficiently accurate values of E, ϕ, and ψ are obtained. The oblate-spheroidal coordinates ρ, σ, and α are then computed; ρ, σ, and α are then used to calculate x, y, and z.

OUTPUT

This program generates position and velocity for equally spaced intervals of time. Oblate-spheroidal coordinates are defined by the equations

$$x = \sqrt{\rho^2 + c^2}\sqrt{1-\sigma^2}\cos\alpha ,$$

$$y = \sqrt{\rho^2 + c^2}\sqrt{1-\sigma^2}\sin\alpha ,$$

$$z = \rho\sigma ,$$

$$r = \sqrt{\rho^2 + c^2(1-\sigma^2)} .$$

The formulas for velocity are given in appendix B, they are

$$\dot{x} = -\dot{\alpha}y + x\left(\frac{\rho\dot{\rho}}{\rho^2+c^2} - \frac{\sigma\dot{\sigma}}{1-\sigma^2}\right) ,$$

$$\dot{y} = +\dot{\alpha}x + y\left(\frac{\rho\dot{\rho}}{\rho^2+c^2} - \frac{\sigma\dot{\sigma}}{1-\sigma^2}\right) ,$$

$$\dot{z} = \rho\dot{\sigma} + \sigma\dot{\rho} ,$$

where

$$\dot{\rho} = \frac{\sqrt{-2\hat{h}}}{\rho^2 + c^2\sigma^2}\, ae\left(\sqrt{(\rho - a\kappa)^2 + a^2\lambda^2}\right)\sin E ,$$

$$\dot{\sigma} = \frac{\sqrt{-2\hat{h}}}{\rho^2 + c^2\sigma^2}\, a\sqrt{1-e^2}\sqrt{\frac{\kappa^2+\lambda^2}{\nu^2}}\sqrt{1 - \ell^2\sin^2\psi}\cos\psi ,$$

$$\dot{\alpha} = \frac{\hat{G}}{(\rho^2 + c^2)(1-\sigma^2)} .$$

REMARKS

The computational procedure as it exists in this report was programmed by the author in single-precision floating-point Fortran for an IBM 7090 computer at the Goddard Space Flight Center. All machine results were compared with hand calculations and the practicality of the method was confirmed. The procedure is presently being compared with both single and double precision numerical integration.

ACKNOWLEDGMENTS

The author wishes to express his appreciation to Dr. John P. Vinti of the National Bureau of Standards, to Mr. Imre G. Izsak of the Smithsonian Astrophysical Observatory, and to Dr. Peter Musen of Goddard Space Flight Center for several valuable discussions during the course of this investigation.

REFERENCES

1. IZSAK, I. G., "A Theory of Satellite Motion About an Oblate Planet, I. A Second-Order Solution of Vinti's Dynamical Problem," Smithsonian Inst., Astrophys. Observ. Res. in Space Science Spec. Rept. No. 52, Nov. 21, 1960.
2. ———, "On Satellite Orbits with Very Small Eccentricities," *Astronom. J.*, **66**(3): 129–131, April 1961.
3. VINTI, J. P., "New Method of Solution for Unretarded Satellite Orbits," *J. Res. Nat. Bur. Standards*, **63**B(2): 105–116, October–December 1959.
4. ———, Theory of an Accurate Intermediary Orbit for Satellite Astronomy," *J. Res. Nat. Bur. Standards*, **65**B(3): 169–201, July–September 1961.
5. HILDEBRAND, F. B., "Introduction to Numerical Analysis," New York: McGraw-Hill, 1956.
6. HOUSEHOLDER, A. S., "Principles of Numerical Analysis," New York: McGraw-Hill, 1953.
7. SCARGOROUGH, J. B., "Numerical Mathematical Analysis," 4th ed., Baltimore: Johns Hopkins Press, 1958.
8. KAULA, W. M., "Recommended Earth Gravitational and Geometrical Constants for Tracking and Orbit Computation," Minutes of Meeting of NASA Earth Model, May 4, 1961, at Marshall Space Flight Center.

Appendix A

LIST OF SYMBOLS

A	coefficient in the quartic polynomial $F(\rho)$. See appendix B.
a	canonical constant, one of Izsak's elements, semimajor axis of the orbit.
a_E	the earth's equatorial radius.
a_0	initial value of the canonical constant a.
B	coefficient in the quartic polynomial $F(\rho)$. See appendix B.
c	a constant defined by Vinti's expression $c^2 = J_2 a_E^2$.
$\hat{c}$	a canonical constant; in the limit as $c^2 \to 0$ of Keplerian motion $\hat{c}$ is the total angular momentum.
E	angle corresponding the eccentric anomaly.
$E(\phi_i, \psi_i)$	the ith value of the eccentric anomaly.
$E(\phi_{i+1}, \psi_{i+1})$	the $(i+1)$th value of the eccentric anomaly.
e	canonical constant, one of Izsak's elements, eccentricity of the orbit.
e_0	initial value of the canonical constant e.
e_*	second eccentricity.
F	the earth's force function.
$F(\rho)$	quartic polynomial fundamental to Vinti's theory.
f_1, f_2, f_3, f_4	representation of a set of four equations to be solved by the Newton–Raphson method.
G	the gravitational constant.
$\hat{G}$	a canonical constant; in the limit as $c^2 \to 0$ of Keplerian motion $\hat{G}$ is the z component of the angular momentum. $\hat{G}$ is positive or negative accordingly as the motion is direct or retrograde.
H	the Hamiltonian.
$H(q_1, q_2, \cdots, q_n;$ $p_1, p_2, \cdots, p_n; t)$	the Hamiltonian function (in which time appears explicitly) of a dynamical system of n degrees of freedom with n generalized coordinates $q_1, q_2, \cdots, q_n$ and the conjugate momenta $p_1, p_2, \cdots, p_n$.
$\hat{h}$	a canonical constant; in the limit as $c^2 \to 0$ of Keplerian motion $\hat{h}$ is the total energy in the orbit and always negative.
I	one of Izsak's elements, inclination of the orbit.
i	angle of inclination.
i_0	initial angle of inclination.
J	the Jacobian matrix of the Newton–Raphson method.
$\lvert J \rvert$	the Jacobian determinant.
J_2	the coefficient of the second-degree Legendre polynomial in the earth's gravitational potential.
K_i	notation used for the coefficients in Kepler equation.
K_0	the value c^2/P_0^2.
k	modulus appearing in elliptic integral of the first kind.
l	modulus appearing in elliptic integral of the first kind.
L	the Lagrangian.
M	the earth's mass.

$\hat{M}$	"mean anomaly".
n_ϕ	the anomalistic mean motion.
$\hat{n}$	a constant used in the generalized Kepler equation, the auxiliary mean motion.
p_0	the value α_2^2/μ.
$P(\rho)$	quartic polynomial fundamental to Vinti's theory.
p_ρ, p_σ, p_α	the generalized momenta.
$Q(\sigma)$	quartic polynomial fundamental to Vinti's theory.
R_i	notation used for the coefficients in the equation for right ascension of the ascending node.
r	the geocentric distance of the satellite.
r_0	the initial geocentric distance of the satellite.
S	canonical constant, one of Izsak's elements, sine of the inclination of the orbit.
T_f	final time.
t	time.
t_0	initial time.
Δt	time interval of integration.
$-\hat{t}$	a canonical constant; in the limit as $c^2 \to 0$ of Keplerian motion $-\hat{t}$ is the time of perigee passage.
U	the speed of the satellite.
V	a "true anomaly" analogous to that in Keplerian motion.
$\hat{V}$	the potential which Vinti obtained in oblate spheroidal coordinates.
v	velocity of the satellite.
v_0	initial velocity of the satellite.
W	"argument of latitude" analogous to that in Keplerian motion.
$\hat{W}$	Hamilton's characteristic function.
X	a trial vector for the solution of a set of nonlinear equations by the Newton–Raphson method.
x, y, z	coordinates in the rectangular system.
x_0, y_0, z_0	the initial values of the coordinates in the rectangular system.
$\dot{x}$, $\dot{y}$, $\dot{z}$	the velocity coordinates in the rectangular system.
$\dot{x}_0$, $\dot{y}_0$, $\dot{z}_0$	the initial value of the velocity coordinates in the rectangular system.
Z_{ij}	element of the inverse Jacobian matrix.
α	the right ascension of the satellite.
α_1, α_2, and α_3	Vinti's canonical constants denoted by Izsak as $\hat{h}$, $\hat{c}$, and $\hat{G}$ respectively.
δ	the declination of the satellite.
ϵ	the motion of perigee.
ϵ	an arbitrarily chosen small positive real number (used as a tolerance in the Newton–Raphson method).
η	the motion of the node.
$\eta_0 \approx \sin I = s$	
κ	a series used in the computation: defined by equation 14.
μ	the product GM where G is the gravitational constant and M is the earth's mass.
ν^2	a dimensionless parameter of the order 10^{-3} in the case of the earth.
ρ, σ, α	coordinates in the oblate spheroidal system.
ρ_0, σ_0, α_0	the initial condition of the coordinates in the oblate spheroidal system.
$\dot{\rho}$, $\dot{\sigma}$, $\dot{\alpha}$	the velocity coordinates in the oblate spheroidal system.
$\dot{\rho}_0$, $\dot{\sigma}_0$, $\dot{\alpha}_0$	the initial conditions of the velocity coordinates in the oblate spheroidal system.
ρ_1	perigee of the orbit.

ρ_2	apogee of the orbit.
ϕ	"true anomaly".
ψ	"argument of latitude".
$\hat{\Omega}$	a canonical constant; in the limit as $c^2 \to 0$ of Keplerian motion $\hat{\Omega}$ is the right ascension of the ascending node.
ω	a canonical constant, one of Izsak's constants, a constant of integration.
$\hat{\omega}$	a canonical constant; in the limit as $c^2 \to 0$ of Keplerian motion $\hat{\omega}$ is the argument of perigee.
$\bar{\omega}$	the mean argument of perigee.

Appendix B

DERIVATION OF THE VELOCITIES IN VINTI'S ACCURATE INTERMEDIARY ORBIT OF AN ARTIFICIAL SATELLITE

INTRODUCTION

Izsak[B1] has given an analytic solution for Vinti's intermediary orbit, with both periodic and secular terms correct through the second order in a certain oblateness parameter $\gamma = c/a$ (to be defined later). His solution giving the position vector of the satellite makes extensive use of Jacobian elliptic functions, linear transformations, and mappings in the complex plane. Vinti[B2] also has given an analytic solution to this problem of satellite motion using rapidly converging infinite series instead of Jacobian elliptic functions. His solution not only gives the periodic terms correct to the second order, but also the secular terms to an arbitrarily high order.

This appendix presents the derivation of the velocity vector through the use of equations from both Vinti and Izsak. However, the orbital elements used in this derivation were introduced by Izsak.

DETERMINATION OF VELOCITY

The oblate spheroidal coordinates ρ, σ, and α are defined by

$$x = \sqrt{\rho^2 + c^2}\ \sqrt{1 - \sigma^2}\ \cos\alpha\ , \tag{B1}$$

$$y = \sqrt{\rho^2 + c^2}\ \sqrt{1 - \sigma^2}\ \sin\alpha\ , \tag{B2}$$

$$z = \rho\sigma\ , \tag{B3}$$

$$r = \sqrt{\rho^2 + c^2(1 - \sigma^2)}\ , \tag{B4}$$

where α is the right ascension of a satellite; r is the geocentric distance; and c is a constant defined by $c^2 = J_2 a_E^2$. The quantity J_2 is the coefficient of the second-degree Legendre polynomial in the earth's gravitational potential

$$V = -\frac{\mu}{r}\left[1 - \sum_{n=1}^{\infty} J_n \left(\frac{a_E}{r}\right)^n P_n(\sin\delta)\right] \tag{B5}$$

where δ is the declination of the satellite, a_E is the equatorial radius of the earth, and $\mu = GM$, where G is the gravitational constant and M the mass of the earth.

Differentiating equations B1–B3 with respect to time we find

$$\dot{x} = -\dot{\alpha}y + x\left(\frac{\rho\dot{\rho}}{\rho^2 + c^2} - \frac{\sigma\dot{\sigma}}{1 - \sigma^2}\right)\ , \tag{B6}$$

$$\dot{y} = +\dot{\alpha}x + y\left(\frac{\rho\dot{\rho}}{\rho^2 + c^2} - \frac{\sigma\dot{\sigma}}{1 - \sigma^2}\right)\ , \tag{B7}$$

$$\dot{z} = \rho\dot{\sigma} + \sigma\dot{\rho}\ . \tag{B8}$$

Squaring and adding equations B6–B8 we obtain

$$U^2 = \dot{x}^2 + \dot{y}^2 + \dot{z}^2$$

$$= \left(\frac{\rho^2 + c^2\sigma^2}{\rho^2 + c^2}\right)\dot{\rho}^2 + \left(\frac{\rho^2 + c^2\sigma^2}{1 - \sigma^2}\right)\dot{\sigma}^2 + (\rho^2 + c^2)(1 - \sigma^2)\dot{\alpha}^2\ . \tag{B9}$$

The expressions for $\dot{\rho}$, $\dot{\sigma}$, and $\dot{\alpha}$ can be obtained from the following equations, which define the generalized momenta:

$$P_\rho = \frac{\partial L}{\partial\dot{\rho}} = h_1^2\dot{\rho} = \frac{\partial S}{\partial\rho} = \pm\frac{\sqrt{P(\rho)}}{\rho^2 + c^2}\ , \tag{B10}$$

$$P_\sigma = \frac{\partial L}{\partial \dot{\sigma}} = h_2^2 \dot{\sigma} = \frac{\partial S}{\partial \sigma} = \pm \frac{\sqrt{Q(\sigma)}}{1 - \sigma^2} , \tag{B11}$$

$$P_\alpha = \frac{\partial L}{\partial \dot{\alpha}} = h_3^2 \dot{\alpha} = \frac{\partial S}{\partial \alpha} = \hat{G}^* , \tag{B12}$$

where

$$h_1^2 = \frac{\rho^2 + c^2\sigma^2}{\rho^2 + c^2} , \quad h_2^2 = \frac{\rho^2 + c^2\sigma^2}{1 - \sigma^2} , \quad h_3^2 = (\rho^2 + c^2)(1 - \sigma^2) , \tag{B13}$$

$$P(\rho) = 2\hat{h}\rho^4 + 2\mu\rho^3 - (\hat{c}^2 - 2c^2\hat{h})\rho^2 + 2c^2\mu\rho - c^2(\hat{c}^2 - \hat{G}^2) , \tag{B14}$$

$$Q(\sigma) = -2c^2\hat{h}\sigma^4 - (\hat{c}^2 - 2c^2\hat{h})\sigma^2 + (\hat{c}^2 - \hat{G}^2) , \tag{B15}$$

$S = S(\rho, \sigma, \alpha)$ is the action function, and L is the Lagrangian given by $L = T - V$, where

$$T = \frac{1}{2}\left(\frac{ds}{dt}\right)^2$$

$$= \frac{1}{2}\left(h_1^2 \dot{\rho}^2 + h_2^2 \dot{\sigma}^2 + h_3^2 \dot{\alpha}^2\right) ,$$

$$V = V(\rho, \sigma) = \frac{-\mu\rho}{\rho^2 + c^2\sigma^2} .$$

Here ds/dt is the speed along the path and $V(\rho, \sigma)$ is the potential function introduced by Vinti.[B3]

The radicand in equation B10 can be written in the form

$$P(\rho) = -2\hat{h}(\rho - \rho_1)(\rho_2 - \rho)(\rho - \rho_3)(\rho - \rho_4) , \tag{B16}$$

where ρ_1, ρ_2, ρ_3, and ρ_4 are the zeros of $P(\rho)$. Izsak[B1] has given the zeros in the form

$$\rho_1 = a(1 - e), \quad \rho_2 = a(1 + e), \quad \rho_3 = a(\kappa - i\lambda), \quad \rho_4 = a(\kappa + i\lambda). \tag{B17}$$

The orbital elements a and e are the semimajor axis and the eccentricity of the orbit, respectively, even though it is not an exact ellipse. They are defined by the first two of equations B17. The quantity i is the imaginary unit $(-1^{1/2})$.

If we substitute ρ_3 and ρ_4 from equations B17 into equation B16:

$$P(\rho) = -2\hat{h}(\rho - \rho_1)(\rho_2 - \rho)\left[\rho^2 - 2a\kappa\rho + a^2(\kappa^2 + \lambda^2)\right] .$$

The quantities κ and $\kappa^2 + \lambda^2$ are given[B1] in terms of a, e, and $s = \sin I$, where I is the inclination of the orbit:

$$\kappa = \frac{\gamma^2(1 - s^2)(1 - e^2 - \gamma^2 s^2)}{(1 - e^2 - \gamma^2)(1 - e^2 - \gamma^2 s^2) + 4\gamma^2 s^2} , \tag{B18}$$

$$\kappa^2 + \lambda^2 = \frac{\gamma^2 s^2\left[(1 - e^2 - \gamma^2)(1 - e^2 - \gamma^2 s^2) + 4\gamma^2\right]}{(1 - e^2 - \gamma^2)(1 - e^2 - \gamma^2 s^2) + 4\gamma^2 s^2} , \tag{B19}$$

where $\gamma = c/a$, a small dimensionless parameter.

*The caret above h, c, and G (that is, $\hat{h}$, $\hat{c}$, and $\hat{G}$) indicates canonical constants, referred to by Izsak (reference B1), where $\hat{h}$ is the total energy in the orbit and always negative, $\hat{c}$ is the total angular momentum, and $\hat{G}$ is the z component of the angular momentum, positive for direct motion.

The quartic $Q(\sigma)$ contains only even powers of σ and can be written

$$Q(\sigma) = -2c^2\hat{h}\left(\sigma_1^2 - \sigma^2\right)\left(\sigma_2^2 - \sigma^2\right) ,$$

where the four real zeros of $Q(\sigma)=0$ are $\pm\sigma_1$ and $\pm\sigma_2$: $0\leq\sigma_1<1$, $\sigma_2\gg1$. As pointed out by Izsak,[B1] σ oscillates between the values $-\sigma_1$ and $+\sigma_1$. Therefore, σ_1 is a convenient parameter to use as the sine of the inclination I of the orbit.

When we introduce Izsak's formulas,

$$\rho_1 = a(1 - e), \quad \rho_2 = a(1 + e) ,$$

$$\rho = a(1 - e\cos E) ,$$

$$\sigma_1 = s = \sin I ,$$

$$\sigma = s\sin\psi ,$$

$$\frac{\sigma_1^2}{\sigma_2^2} = \frac{s^2}{\sigma_2^2} = l^2 ,$$

$$c^4 s^2 \frac{s^2}{l^2} = a^4\left(1 - e^2\right)\left(\kappa^2 + \lambda^2\right) ,$$

where E is the eccentric anomaly and ψ is the argument of latitude, and several of the aforementioned relations into equations B10 and B11 we obtain

$$P_\rho = \frac{\sqrt{-2\hat{h}}}{\rho^2 + c^2} ae \sqrt{\rho^2 - 2a\kappa\rho + a^2\left(\kappa^2 + \lambda^2\right)} \sin E , \tag{B20}$$

$$P_\sigma = \frac{\sqrt{-2\hat{h}}}{1 - \sigma^2}\frac{a^2}{c}\sqrt{1 - e^2}\sqrt{\kappa^2 + \lambda^2}\sqrt{1 - l^2\sin^2\psi}\cos\psi . \tag{B21}$$

The coefficients and the roots of the quartic equation $P(\sigma)=0$ can be related to those of $Q(\sigma)=0$ in the following manner:

$$2a(1 + \kappa) = -\frac{\mu}{\hat{h}} , \tag{B22}$$

$$a^2\left[\left(1 - e^2\right) + 4\kappa + \left(\kappa^2 + \lambda^2\right)\right] = -\frac{\hat{\mathcal{E}}^2 - 2c^2\hat{h}}{2\hat{h}}$$

$$= c^2\left(s^2 + \frac{s^2}{l^2}\right) , \tag{B23}$$

$$2a^3\left[(1-e^2)\kappa+(\kappa^2+\lambda^2)\right] = -c^2\frac{\mu}{\hat{h}} \quad , \tag{B24}$$

$$a^4(1-e^2)(\kappa^2+\lambda^2) = -c^2\frac{\hat{c}^2-\hat{G}^2}{2\hat{h}}$$

$$= c^4s^2\frac{s^2}{l^2} \quad . \tag{B25}$$

Consider the following expression for l^2 given by Izsak[B1]

$$l^2 = \frac{\gamma^2s^2}{1-e^2}\left[\frac{(1-e^2-\gamma^2)(1-e^2-\gamma^2s^2)+4\gamma^2s^2}{(1-e^2-\gamma^2)(1-e^2-\gamma^2s^2)+4\gamma^2}\right] . \tag{B26}$$

If we substitute c/a for γ and solve for s^2/l^2 we obtain

$$\frac{s^2}{l^2} = \frac{a^2(1-e^2)}{c^2}\left\{\frac{\left[a^2(1-e^2)-c^2\right]\left[a^2(1-e^2)-c^2s^2\right]+4a^2c^2}{\left[a^2(1-e^2)-c^2\right]\left[a^2(1-e^2)-c^2s^2\right]+4a^2c^2s^2}\right\} . \tag{B27}$$

Next we introduce a parameter $p=a(1-e^2)$, the semilatus rectum, which Vinti[B2] uses in his oblateness parameter $k=c^2/p^2$. It is clear that equation B27 can be written

$$\frac{s^2}{l^2} = \frac{ap}{c^2}\left[\frac{(ap-c^2)(ap-c^2s^2)+4a^2c^2}{(ap-c^2)(ap-c^2s^2)+4a^2c^2s^2}\right] . \tag{B28}$$

From equation 4.13 of reference B2:

$$B = c^2\eta_0^2\,\frac{(ap-c^2)(ap-c^2\eta_0^2)+4a^2c^2}{(ap-c^2)(ap-c^2\eta_0^2)+4a^2c^2\eta_0^2} , \tag{B29}$$

where $\eta_0=s=\sin I$, we see that $c^4s^4/l^2=Bap$. Solving for c^2s^2/l^2 from equation B23 and inserting it into equation B25, we obtain

$$\hat{G}^2 = (1-s^2)(\hat{c}^2+2c^2\hat{h}s^2) . \tag{B30}$$

From equation B23,

$$\frac{-\hat{c}^2}{2\hat{h}} = -c^2(1-s^2)+c^2\frac{s^2}{l^2} . \tag{B31}$$

Multiplying equation B28 by c^2 to obtain c^2s^2/l^2 and inserting it into equation B31 we find

$$\frac{-\hat{c}^2}{2\hat{h}} = -c^2(1-s^2)+ap\left[\frac{(ap-c^2)(ap-c^2s^2)+4a^2c^2}{(ap-c^2)(ap-c^2s^2)+4a^2c^2s^2}\right] . \tag{B32}$$

Vinti has given an identical result in equation 4.15 of reference B2:

$$\frac{-a_2^2}{2a_1} \equiv a_0p_0 = -c^2(1-\eta_0^2)+ap\left[\frac{(ap-c^2)(ap-c^2\eta_0^2)}{(ap-c^2)(ap-c^2\eta_0^2)}\,\frac{+4a^2c^2}{+4a^2c^2\eta_0^2}\right] , \tag{B33}$$

where a_0 is a semimajor axis, p_0 the semilatus rectum. Since $\hat{h}=\alpha_1$, $\hat{c}=\alpha_2$, and $\hat{G}=\alpha_3$, we can easily rewrite equation B30, using equation B33, to obtain the final result for $\hat{G}$,

$$\hat{G} = \hat{c}\sqrt{\left(1-\frac{c^2s^2}{a_0p_0}\right)(1-s^2)} . \tag{B34}$$

Equation B34 is equivalent to Vinti's equation 4.15a of reference B2:

$$\alpha_3 = \alpha_2 \left(1 - \frac{c^2\eta_0^2}{a_0 p_0}\right)^{\frac{1}{2}} \cos I.$$

Using equation B34 we obtain, as in equation B12,

$$P_a = \hat{G}. \tag{B35}$$

It should be noted that the following formulas relate κ and $\kappa^2+\lambda^2$ to Vinti's A and B:

$$a^2(\kappa^2 + \lambda^2) = B \quad , \tag{B36}$$

$$-2a\kappa = A, \tag{B37}$$

where A and B are given by equations 4.12 and 4.13 of reference B2:

$$\left.\begin{aligned} A &= \frac{-2ac^2(1-\eta_0^2)(ap - c^2\eta_0^2)}{(ap - c^2)(ap - c^2\eta_0^2) + 4a^2c^2\eta_0^2}, \\ B &= c^2\eta_0^2 \frac{(ap - c^2)(ap - c^2\eta_0^2) + 4a^2c^2}{(ap - c^2)(ap - c^2\eta_0^2) + 4a^2c^2\eta_0^2}. \end{aligned}\right\} \tag{B38}$$

From Vinti's equation 4.16 (reference B2) we have

$$\eta_2^{-2} = \frac{c^2}{ap} \frac{(ap - c^2)(ap - c^2\eta_0^2) + 4a^2c^2\eta_0^2}{(ap - c^2)(ap - c^2\eta_0^2) + 4a^2c^2} = c^4\eta_0^2 (Bap)^{-1}, \tag{B39}$$

where $\eta_2 = \sigma_2$. Using equations B36–B38, together with the values for $P(\rho)$ and $Q(\sigma)$, we find that in Vinti's notation the quartics $P(\rho)$ and $Q(\sigma)$ can be factored in the form

$$F(\rho) = -2\alpha_1 (\rho_2 - \rho)(\rho - \rho_1)(\rho^2 + A\rho + B), \tag{B40}$$

$$G(\eta) = -2\alpha_1 c^2 (\eta_0^2 - \eta^2)(\eta_2^2 - \eta^2), \tag{B41}$$

where $\eta = \sigma$.

The following equations for $\dot{\rho}$, $\dot{\sigma}$, and $\dot{\alpha}$ are easily obtained from equations B10–B13, B20, B21, and B35:

$$\dot{\rho} = \frac{\sqrt{-2h}}{\rho^2 + c^2\sigma^2} ae \sqrt{\rho^2 - 2a\kappa\rho + a^2(\kappa^2 + \lambda^2)} \sin E, \tag{B42}$$

$$\dot{\sigma} = \frac{\sqrt{-2h}}{\rho^2 + c^2\sigma^2} \frac{a^2}{c} \sqrt{1 - e^2} \sqrt{\kappa^2 + \lambda^2} \sqrt{1 - l^2 \sin^2\psi} \cos\psi, \tag{B43}$$

$$\dot{\alpha} = \frac{\hat{G}}{(\rho^2 + c^2)(1 - \sigma^2)}. \tag{B44}$$

If we write the equations for $\dot{\rho}$, $\dot{\sigma}$, and $\dot{\alpha}$ in Vinti's notation, we obtain

$$\dot{\rho} = \frac{\sqrt{-2\alpha_1}}{\rho^2 + c^2\eta^2} \, ae \, \sqrt{\rho^2 + A\rho + B} \, \sin E, \tag{B45}$$

$$\dot{\eta} = \frac{\sqrt{-2\alpha_1}}{\rho^2 + c^2\eta^2} \, c\eta_0\eta_2 \, \sqrt{1 - q^2 \sin^2\psi} \, \cos\psi, \tag{B46}$$

$$\dot{\phi} = \frac{\alpha_3}{(\rho^2 + c^2)(1 - \eta^2)}, \tag{B47}$$

where $q = \eta_0/\eta_2$.

Now, substituting equations B42–B44 into equations B6–B8 we obtain the following:

$$\dot{x} = \frac{-\hat{G}y}{(\rho^2 + c^2)(1 - \sigma^2)} + \frac{xa\sqrt{-2\hat{h}}}{\rho^2 + c^2\sigma^2} \left[\frac{\rho e \sqrt{\rho^2 - 2a\kappa\rho + a^2(\kappa^2 + \lambda^2)}}{\rho^2 + c^2} \sin E \right.$$

$$\left. - \frac{\sigma\sqrt{1 - e^2}}{1 - \sigma^2} \frac{a}{c} \sqrt{\kappa^2 + \lambda^2} \sqrt{1 - l^2 \sin^2\psi} \cos\psi \right], \tag{B48}$$

$$\dot{y} = \frac{+\hat{G}x}{(\rho^2 + c^2)(1 - \sigma^2)} + \frac{ya\sqrt{-2\hat{h}}}{\rho^2 + c^2\sigma^2} \left[\frac{\rho e \sqrt{\rho^2 - 2a\kappa\rho + a^2(\kappa^2 + \lambda^2)}}{\rho^2 + c^2} \sin E \right.$$

$$\left. - \frac{\sigma\sqrt{1 - e^2}}{1 - \sigma^2} \frac{a}{c} \sqrt{\kappa^2 + \lambda^2} \sqrt{1 - l^2 \sin^2\psi} \cos\psi \right], \tag{B49}$$

$$\dot{z} = \frac{a\sqrt{-2\hat{h}}}{\rho^2 + c^2\sigma^2} \left[\frac{a}{c} \rho \sqrt{1 - e^2} \sqrt{\kappa^2 + \lambda^2} \sqrt{1 - l^2 \sin^2\psi} \cos\psi \right.$$

$$\left. + \sigma e \sqrt{\rho^2 - 2a\kappa\rho + a^2(\kappa^2 + \lambda^2)} \sin E \right]. \tag{B50}$$

The velocity components given above are now being used in an orbit determination program formulated by the author.

REFERENCES

B1. Izsak, I. G., "A Theory of Satellite Motion About an Oblate Planet. I. A Second-Order Solution of Vinti's Dynamical Problem," Smithsonian Inst., Astrophys. Observ. Res. in Space Science Spec. Rept. No. 52, Nov. 21, 1960.

B2. Vinti, J. P., "Theory of an Accurate Intermediary Orbit for Satellite Astronomy," *J. Res. Nat. Bur. Standards*, **65**B(3): 169–201, July–September 1961.

B3. ———, "New Method of Solution for Unretarded Satellite Orbits," *J. Res. Nat. Bur. Standards*, **63**B(2): 105–116, October–December 1959.

COMPUTATION OF GENERAL PLANETARY PERTURBATIONS, PART I.

LLOYD CARPENTER

Goddard Space Flight Center

A computer program for automatic computation of first order general planetary perturbations is described. The program is based on Hansen's theory as given in the Auseinandersetzung. As examples the general perturbations of six minor planets are given.

INTRODUCTION

The need for general perturbations of minor planets, comets, and artificial satellites has been widely recognized, and these series developments remain of interest despite well developed techniques for numerical integration on electronic computers. For purposes of long range predictions and studies of stability of orbits with peculiar elements, the perturbations in series are of particular importance. As artificial satellites are launched into orbits with larger semi-major axes, the study of their behavior under the influence of the moon will bear close resemblance to the planetary or cometary problem. Advances in the speed and capabilities of electronic computers have reduced the programming of general perturbations to a reasonable task.

With these facts in mind a program was developed for automatic machine computation of general planetary perturbations using Hansen's theory as given in the Auseinandersetzung (1857).[1] Hill's modification for computing the perturbations in the radius vector[2] and Herrick's suggested use of Gibbs' vectorial elements have been included. Hansen's method was chosen because it can be applied to a wide range of eccentricities and inclinations. No exposition of the theory will be given since no modifications have been made. The details involved in the computational procedure are presented in appendix A. Perturbations for several minor planets are given in appendix B.

GENERAL DISCUSSION OF COMPUTATIONS

The computational procedure was taken from Herget[3] except that the Laplace coefficients are computed directly. The program takes the elements of the disturbed and disturbing bodies as the input data and prints out the coefficients in the series for the perturbations of first order as the results. The time required on the IBM 7094 for each set of perturbations is approximately 1 minute per planet.

The following set of equations is used in developing the series for the perturbations (reference 3, equations 8, 6):*

$$3\,a\,\Omega = 3\,m'\,a\,\Delta^{-1} + (-3\mathfrak{K}) ,$$

$$a\,r\,\frac{\partial \Omega}{\partial r} = m'\,a\,\Delta^{-3}\left(\frac{r'^2 - r^2}{2}\right) - \frac{1}{6}\,(3\,a\,\Omega) + \frac{1}{2}\,(-3\mathfrak{K})$$

*See appendix A for notation used and computational details.

The disturbed position vector is then given by

$$\mathbf{r} = (1+\nu)\left[\mathbf{A}(\cos\bar{E}-e) + \mathbf{B}\sin\bar{E} + \mathbf{C}u\right] .$$

The velocity vector may also be determined, but with less accuracy, by evaluating the derivatives of the series for the perturbations. Thus for any given time the osculating elements of the disturbed motion may be obtained. Taking $\mathbf{v}=d\mathbf{r}/d\tau$ where $\tau=k(t-t_0)$ and k is the Gaussian constant, we have

$$\mathbf{v} = \left[\mathbf{A}(\cos\bar{E}-e) + \mathbf{B}\sin\bar{E} + \mathbf{C}u\right]\frac{d\nu}{d\tau}$$

$$+ (1+\nu)\left\{\left[-\mathbf{A}\sin\bar{E}+\mathbf{B}\cos\bar{E}\right]\frac{d\bar{E}}{d\tau} + \mathbf{C}\frac{du}{d\tau}\right\},$$

where

$$\frac{d\bar{E}}{d\tau} = \frac{a^{-3/2} + \frac{d}{d\tau}(n\delta z)}{1 - e\cos\bar{E}} ;$$

and the derivatives of the perturbations are computed by using the relation

$$\frac{d}{d\tau} = \frac{1}{r\sqrt{a}}\frac{d}{d\bar{E}} .$$

The example given by Herget[3] of the perturbations of (1286) Banachiewizca by Jupiter was used to check all the intermediate results during the programing process. After the program was completed, a comparison was made with the first order perturbations of (13) Egeria computed by Hansen and given in his original work.[1] The agreement in the perturbations due to Jupiter and Saturn in this case were most encouraging, the largest difference between corresponding periodic terms was less than 1 second of arc, while for the secular and mixed terms the differences were of the order of 10^{-4} seconds of arc or less. For the small perturbations due to Mars there was some disagreement. These comparisons were made without including the constants of integration, since Hansen computed these terms once including the perturbations due to Jupiter, Saturn, and Mars together. Additional comparisons were made with various perturbations which have been computed using the same method. The elements given by the original authors were used for these comparisons. The examples of perturbations given in appendix B are based on the elements of minor planets given in reference 4 and the elements of Jupiter given by Clemence.[5] The main differences in the perturbations computed using different sets of elements appear in the terms affected by the constants of integration and in the long period terms associated with small divisors.

CONCLUSION

We now have the facility for automatic computation of Hansen's first order planetary perturbations. These perturbations are sufficiently accurate for the practical purposes of identification and producing ephemerides in the case of planetary-type motion.

Several authors have made contributions to the development into Fourier series of lunar perturbations of artificial satellites. The works of Kozai[6]; Musen, Bailie, and Upton[7]; and Kaula[8] should be mentioned. The analytic development in powers of the ratio of the semi-major axes converges rapidly for close earth satellites, but for more distant satellites the convergence is slow. For the latter cases one must apply harmonic analysis as in Hansen's planetary theory. For this reason the program described in this article is now the most efficient method of treating periodic perturbations of cislunar satellites.

The continuation of this work has several different aspects. A problem of considerable interest and importance is increasing the range of applicability so that general perturbations of highly eccentric orbits can be computed. A modification of the present program which accomplishes this purpose by direct double harmonic analysis will be described in part II. A corresponding program is planned using the mean anomaly as the independent variable for greater convenience in evaluating the perturbations. For greater accuracy it is desirable to have a program for the computation of perturbations of higher order. For this purpose the equations in rectangular coordinates given by Musen and Carpenter[9] have a convenient form for programing. First order perturbations in rectangular coordinates will be compared with Hansen's perturbations. It also would be desirable to have a program using mean

$$a^2 \frac{\partial \Omega}{\partial Z} = m' a\left(\Delta^{-3} - r'^{-3}\right) Z' ,$$

$$W = \int\left[\frac{1}{3} M \frac{\partial(3a\Omega)}{\partial E} + N a r \frac{\partial \Omega}{\partial r}\right] dE ,$$

$$R = \int Q a^2 \frac{\partial \Omega}{\partial Z} \, dE ,$$

$$n\delta z = \int \bar{W}(1 - e\cos E)\, dE ,$$

$$\nu = -\frac{1}{6} X_0 - \frac{e}{6} X_1 - \frac{1}{2} \bar{W} ,$$

$$u = \bar{R} .$$

A major part of the computation is spent in the development of Δ^{-1} and Δ^{-3} where Δ is the distance between the disturbed and disturbing bodies. These quantities are expanded into double Fourier series in terms of the eccentric anomalies by taking the first 16 Laplace coefficients and applying harmonic analysis with 24 equally spaced values of the eccentric anomaly of the disturbed body. Representing the argument of any term in the form $(iE-jE')$, the terms which are computed correspond to $j=0$ through 15 and $i=j-11$ through $j+11$. The derivatives of the disturbing function

$$\partial(3a\Omega)/\partial E, \quad ar(\partial\Omega/\partial r), \quad \text{and} \quad a^2(\partial\Omega/\partial Z)$$

are obtained by simple operations on the series Δ^{-1} and Δ^{-3}. These series are then transformed to arguments of the form $(iE-j\phi)$ where

$$\phi = (n'/n)(E-g_0)+g_0'$$

by applying the Bessel transformations. The multiplications by the M, N, and Q expressions which are given in the Collection of Formulas are combined with the term by term integrations by forming sums of products of coefficients and dividing by $i-jn'/n$. The constant terms yield terms factored by E after the integration, and these are converted to coefficients of time by replacing E by $nt + e\sin E$. The replacement bar operation is accomplished by considering the temporary angle H to be the same as E and combining corresponding coefficients. After the formal integration is completed, the constants of integration are determined so as to satisfy the initial conditions.

The perturbations are used in the following manner. For any given time, solve the Kepler's equation

$$E - e\sin E = g_0 + n(t - t_0)$$

to obtain the undisturbed eccentric anomaly E. E must not be reduced modulus 360° because it appears in the noninteger multiples through ϕ. With this value of E the argument for each term is computed:

$$iE - j\phi = \left(i - j\frac{n'}{n}\right)E + j\left(\frac{n'}{n} g_0 - g_0'\right) .$$

Evaluate the series for $n\delta z$, ν and u by multiplying the coefficients by the cosine or sine of $iE-j\phi$ and adding the terms of the series. With the value of $n\delta z$ Kepler's equation is solved for the disturbed eccentric anomaly

$$\bar{E} - e\sin\bar{E} = g_0 + n(t - t_0) + n\delta z .$$

elements and a program for orbit correction based on general perturbations.

ACKNOWLEDGMENTS

The author expresses his gratitude to Dr. Peter Musen, who suggested the work on general planetary perturbations and gave generously of his time in many helpful discussions. Thanks are also given to Neal Paris, who assisted in modifying the program to include direct double harmonic analysis.

REFERENCES

1. HANSEN, P. A., "Auseinandersetzung einer zweck-Massigen methode zur Berechnung der Absoluten Störungen der Kleinen Planeten," Leipzig, 1857–59.
2. HILL, G. W., "The Collected Works of George William Hill," Vol. I, Washington, D.C.: Carnegie Institution, 1905, p. 348.
3. HERGET, P. H., "The Computation of Orbits," 1948.
4. "Ephemerides of Minor Planets for 1962," Institute of Theoretical Astronomy, Academy Nauk U.S.S.R., 1961.
5. CLEMENCE, G. M., "On the Elements of Jupiter," *Astronom. J.*, **52**(1160): 89–93, November 1946.
6. KOZAI, Y., "On the Effects of the Sun and the Moon Upon the Motion of a Close Earth Satellite," Smithsonian Inst. Astrophys. Obs. Spec. Rept. No. 22, 1959, pp. 7–10.
7. MUSEN, P., BAILIE, A., and UPTON, E., "Development of the Lunar and Solar Perturbations in the Motion of an Artificial Satellite," NASA Technical Note D–494, 1961.
8. KAULA, W. M., "A Development of the Lunar and Solar Disturbing Functions for a Close Satellite," *Astronom. J.*, **67**, p. 300 (1962).
9. MUSEN, P., and CARPENTER, L., "On the General Planetary Perturbations in Rectangular Coordinates," *J. Geophys. Res.*, May 1963.

Appendix A

COLLECTION OF FORMULAS

NOTATION

$g_0, a, e, n, \omega, \Omega, i,$	the osculating elements of the disturbed body.
$g_0', a', e', n', \omega', \Omega', i'$	the orbital elements of the disturbing body.
Ω	the disturbing function.
$\mathbf{r}$	the position vector of the disturbed body.
$\mathbf{v}$	the velocity vector of the disturbed body.
$r = \lvert \mathbf{r} \rvert$	
$\mathbf{r}'$	the position vector of the disturbing body.
$r' = \lvert \mathbf{r} \rvert$	
Δ	the mutual distance $\Delta^2 = r^2 + r'^2 - 2\mathbf{r}\cdot\mathbf{r}'$.
m'	the mass of the disturbing body.
$\mathfrak{C} = m'a\mathbf{r}\cdot\mathbf{r}'/r'^3$.	
H	the fictitious eccentric anomaly to be replaced by E after the integration.
E	the eccentric anomaly of the disturbed body.
$\mathbf{A}$	the vector of length a in the direction of perigee of the undisturbed orbit.
$\mathbf{B}$	the vector of length $a(1-e^2)^{1/2}$ in the direction 90° in advance of perigee of the undisturbed orbit.
$\mathbf{C}$	the vector of length a in the direction of the angular momentum of the undisturbed orbit; the vectors **A**, **B**, and **C** are referred to the equatorial system of coordinates.
Z	the coordinate in the direction of **C**.
$n\delta z$	the perturbation in the mean anomaly.
ν	the perturbation in the length of the radius vector.
u	the perturbation normal to the orbit plane.
$\phi = n'/n(E-g_0)+g_0'$.	
E_0	the eccentric anomaly of the disturbed body at the epoch.
ϵ	the obliquity of the ecliptic.
$b_s^{(j)}$	the Laplace coefficients.

For purposes of computation, the remaining symbols used in the Collection of Formulas are defined by the expressions in which they appear.

THE PROCEDURE

1. Take the osculating elements

$$\mathbf{a},\ \mathbf{e},\ \mathbf{i},\ \omega,\ \Omega,\ \mathbf{g}_0,\ \mathbf{n}$$

of the disturbed planet and the elements

$$\mathbf{a}',\ \mathbf{e}',\ \mathbf{i}',\ \omega',\ \Omega',\ \mathbf{g}_0',\ \mathbf{n}',\ \mathbf{m}'$$

of the disturbing planet as input data.

2. Evaluate the vectors

$$\mathbf{A} = a\,G(\epsilon) \cdot \begin{pmatrix} \cos\omega\cos\Omega - \sin\omega\sin\Omega\cos i \\ \cos\omega\sin\Omega + \sin\omega\cos\Omega\cos i \\ \sin\omega\sin i \end{pmatrix},$$

$$\mathbf{B} = a\sqrt{1-e^2}\,G(\epsilon) \cdot \begin{pmatrix} -\sin\omega\cos\Omega - \cos\omega\sin\Omega\cos i \\ -\sin\omega\sin\Omega + \cos\omega\cos\Omega\cos i \\ \cos\omega\sin i \end{pmatrix},$$

$$\mathbf{C} = a\,G(\epsilon) \cdot \begin{pmatrix} \sin\Omega\sin i \\ -\cos\Omega\sin i \\ \cos i \end{pmatrix},$$

and the corresponding primed vectors for the disturbing planet. The rotation matrix $G(\epsilon)$ refe the vectors to the equatorial system of coordinates:

$$G(\epsilon) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\epsilon & -\sin\epsilon \\ 0 & \sin\epsilon & \cos\epsilon \end{pmatrix}.$$

3. Compute the dot products

$$\mathbf{A}\cdot\mathbf{A}',\ \mathbf{A}\cdot\mathbf{B}',\ \mathbf{B}\cdot\mathbf{A}',\ \mathbf{B}\cdot\mathbf{B}',\ \mathbf{C}\cdot\mathbf{A}',\ \mathbf{C}\cdot\mathbf{B}'.$$

4. Carry out steps 4a through 4e for each of the following values of the eccentric anomaly E of the disturbed planet: $E=0°, 15°, 30°, \cdots, 345°$

   a. Compute the following quantities:

$$r = a(1 - e\cos E)$$

$$K\cos\psi = 2e'a'^2 - 2e\mathbf{A}\cdot\mathbf{A}' + 2\mathbf{A}\cdot\mathbf{A}'\cos E + 2\mathbf{B}\cdot\mathbf{A}'\sin E$$

$$K\sin\psi = -2e\mathbf{A}\cdot\mathbf{B}' + 2\mathbf{A}\cdot\mathbf{B}'\cos E + 2\mathbf{B}\cdot\mathbf{B}'\sin E$$

$$H = a'^2\left(1 - 2e'^2\right) + r^2 + e'(K\cos\psi).$$

b. Solve by iteration for C, q, and Q starting with $Cw/q^2=0$, where $w=a'^2e'^2$:

$$q\cos Q = \frac{K\cos\psi}{1 + Cw/q^2}$$

$$q\sin Q = \frac{K\sin\psi}{1 - Cw/q^2}$$

$$C = H + \frac{w}{q^2}(q\sin Q)^2 .$$

c. Compute the following quantities:

$$A = \frac{q}{C + \sqrt{C^2 - q^2}}$$

$$Q' = Q - E$$

$$1.5\frac{w}{q}\cos Q'$$

$$1.5\frac{w}{q}\sin Q'$$

$$\frac{15}{16}\left(\frac{w}{q}\right)^2$$

$$\frac{15}{16}\left(\frac{w}{q}\right)^2\cos 2Q'$$

$$\frac{15}{16}\left(\frac{w}{q}\right)^2\sin 2Q'$$

d. Compute the Laplace coefficients $b_s^{(j)}$ for $s=\frac{1}{2}$ and $\frac{3}{2}$ and $j=0$ through 15 as follows:

i. Compute $b_{1/2}^{(14)}$ and $b_{1/2}^{(15)}$ by numerical evaluation of the integral

$$b_{1/2}^{(j)} = \frac{4}{\pi}A^j\int_0^{\pi/2}\frac{\sin^{2j}\theta}{\sqrt{1 - A^2\sin^2\theta}}d\theta .$$

ii. Compute $b_{1/2}^{(j)}$ for $j=13, 12, \cdots, 0$ by the recurrence relation

$$b_{1/2}^{(j)} = \frac{2j+2}{2j+1}\left(A + \frac{1}{A}\right)b_{1/2}^{(j+1)} - \frac{2j+3}{2j+1}b_{1/2}^{(j+2)} .$$

iii. Compute $b_{3/2}^{(15)}$ from the equation

$$b_{3/2}^{(15)} = -29\frac{1 + A^2}{(1 - A^2)^2}b_{1/2}^{(15)} + 58\frac{A}{(1 - A^2)^2}b_{1/2}^{(14)} .$$

iv. Compute $b_{3/2}^{(j)}$ for $j=0, 1, \cdots, 14$ from the recurrence relation

$$b_{3/2}^{(j)} = (2j+1)\frac{(1+A^2)}{(1-A^2)^2} b_{1/2}^{(j)} - 2(2j+1)\frac{A}{(1-A^2)^2} b_{1/2}^{(j+1)} .$$

e. By taking $k=q/2A$, compute the following for $s=\frac{1}{2}$ and $\frac{3}{2}$ and $j=0$ through 15

$$k^{-s} b_s^{(j)} \cos j Q'$$

$$k^{-s} b_s^{(j)} \sin j Q' .$$

5. Compute the coefficients $C_{j,h}^*$, $S_{j,h}^*$, $C_{j,h}^{*\prime}$, and $S_{j,h}^{*\prime}$ in the following expansions by harmonic analysis of the tabulated values:

$$k^{-s} b_s^{(j)} \cos j Q' = \sum_h \left[C_{j,h}^* \cos (hE) + S_{j,h}^* \sin (hE)\right]$$

$$k^{-s} b_s^{(j)} \sin j Q' = \sum_h \left[C_{j,h}^{*\prime} \cos (hE) + S_{j,h}^{*\prime} \sin (hE)\right] ,$$

for $s=\frac{1}{2}$ and $\frac{3}{2}$; $j=0$ through 15 and $h=0$ through 11.

6. Compute the coefficients c_{ij} and s_{ij} in the following expansions

$$\left[C - q\cos (E-E'+Q')\right]^{-s} = \sum_i \sum_j \left[c_{ij} \cos (iE - jE') + s_{ij} \sin (iE - jE')\right]$$

for $s=\frac{1}{2}$ and $\frac{3}{2}$; $j=0$ through 15 and $i=j-11$ through $j+11$ using the formula

$$\left[C - q\cos (E-E'+Q')\right]^{-s} = \sum_j \left[\left(k^{-s} b_s^{(j)} \cos j Q'\right) \cos j (E-E')\right.$$

$$\left. -\left(k^{-s} b_s^{(j)} \sin j Q'\right) \sin j (E-E')\right] .$$

7. Evaluate the coefficients C_{ij} and S_{ij} in the series

$$\left[1 - \frac{w}{q}\cos (E+E'+Q')\right]^{-s} = 1 + \sum_{i,j} \left[C_{i,j} \cos (iE - jE') + S_{i,j} \sin (iE - jE')\right]$$

for $s=\frac{1}{2}$ and $\frac{3}{2}$, and $j=0$, 1, and 2 by harmonic analysis over E of the coefficients in the first three terms of the binomial expansion of the left-hand side.

8. Obtain double Fourier series for Δ^{-1} and Δ^{-3} from the product

$$\Delta^{-2s} = \left[C - q\cos (E-E'+Q')\right]^{-s} \cdot \left[1 - \frac{w}{q}\cos (E+E'+Q')\right]^{-s} .$$

9. Expand $(a'/r')^3$ in a cosine series in E' by harmonic analysis of the expression

$$\left(\frac{a'}{r'}\right)^3 = \frac{1}{(1-e'\cos E')^3} .$$

10. Evaluate the coefficients in the expression

$$(2\mathbf{r}\cdot\mathbf{r}') = 2e\,e'\,\mathbf{A}\cdot\mathbf{A}'$$

$$-2e'\,\mathbf{A}\cdot\mathbf{A}'\cos E - 2e'\,\mathbf{B}\cdot\mathbf{A}'\sin E$$

$$+\left[\mathbf{A}\cdot\mathbf{A}'-\mathbf{B}\cdot\mathbf{B}'\right]\cos(-E-E') - \left[\mathbf{B}\cdot\mathbf{A}'+\mathbf{A}\cdot\mathbf{B}'\right]\sin(-E-E')$$

$$-2e\,\mathbf{A}\cdot\mathbf{A}'\cos(-E') + 2e\,\mathbf{A}\cdot\mathbf{B}'\sin(-E')$$

$$+\left[\mathbf{A}\cdot\mathbf{A}'+\mathbf{B}\cdot\mathbf{B}'\right]\cos(E-E') + \left[\mathbf{B}\cdot\mathbf{A}'-\mathbf{A}\cdot\mathbf{B}'\right]\sin(E-E') .$$

11. Evaluate the coefficients in the series

$$-3\mathcal{K} = -\frac{3m'\,a}{2a'^3}\left(\frac{a'}{r'}\right)^3(2\mathbf{r}\cdot\mathbf{r}') .$$

12. Evaluate the coefficients in the series

$$3a\Omega = 3m'\,a\Delta^{-1} + (-3\mathcal{K}) .$$

13. Evaluate the coefficients in the series $\partial(3a\Omega)/\partial E$ from those of $3a\Omega$ by differentiating term by term.

14. Evaluate the coefficients in the expression

$$\left(r'^2 - r^2\right) = a'^2 - a^2 + \frac{a'^2e'^2}{2} - \frac{a^2e^2}{2}$$

$$-2a'^2e'\cos(-E') + 2a^2e\cos E$$

$$+\frac{a'^2e'^2}{2}\cos(-2E') - \frac{a^2e^2}{2}\cos 2E .$$

15. Evaluate the coefficients in the series

$$ar\,\frac{\partial\Omega}{\partial r} = \frac{1}{2}m'\,a\left(r'^2 - r^2\right)\Delta^{-3} - \frac{1}{6}(3a\Omega) + \frac{1}{2}(-3\mathcal{K}) .$$

16. Evaluate the coefficients in the expression

$$Z' = -\mathbf{C}\cdot\mathbf{A}'e' + \mathbf{C}\cdot\mathbf{A}'\cos(-E') - \mathbf{C}\cdot\mathbf{B}'\sin(-E').$$

17. Evaluate the coefficients in the series

$$a^2(\partial\Omega/\partial Z) = m'a(\Delta^{-3} - r'^{-3})Z'.$$

18. Compute the Bessel function coefficients:

$$P_k^{(j)} = \frac{j}{k}J_{k-j}(ke') \quad \text{for} \quad k, j = 1,2,3,\cdots,15,$$

$$P_0^{(1)} = -\frac{e'}{2} ,$$

$$P_0^{(j)} = 0 \quad \text{for} \quad j \neq 1 ,$$

where

$$J_m(x) = \sum_{l=0} \frac{(-1)^l \left(\frac{x}{2}\right)^{2l+m}}{l!\,(l+m)!}$$

and

$$J_{-m}(x) = (-1)^m J_m(x) .$$

19. Transform the series for $\partial(3a\Omega)/\partial E$, $ar(\partial\Omega/\partial r)$, and $a^2(\partial\Omega/\partial Z)$ from arguments of the form $iE-jE'$ to arguments of the form $iE-jg'$ using the expansion

$$\begin{matrix}\cos\\ \sin\end{matrix}(iE-jE') = \sum_{k=-\infty}^{+\infty} P_k^{(j)} \begin{matrix}\cos\\ \sin\end{matrix}(iE-kg')$$

for each term.

20. Transform the series for $\partial(3a\Omega)/\partial E$, $ar(\partial\Omega/\partial r)$, and $a^2(\partial\Omega/\partial Z)$ from arguments $iE-jg'$ to arguments $iE-j\phi$, where $\phi=(n'/n)E-(n'/n)g_0+g_0'$, using the expansion

$$\begin{matrix}\cos\\ \sin\end{matrix}(iE-jg') = \sum_{k=-\infty}^{+\infty} J_{k-i}\left(je\frac{n'}{n}\right) \begin{matrix}\cos\\ \sin\end{matrix}(kE-j\phi) ,$$

where the Bessel functions are computed as in step 18.

21. Evaluate the coefficients in the expression

$$Q = e\sin(E) - \frac{1}{2}e^2\sin(2E) + \frac{1}{2}e^2\sin(H+E)$$

$$-\frac{3}{2}e\sin(H) + \left(1+\frac{1}{2}e^2\right)\sin(H-E) - \frac{1}{2}e\sin(H-2E) .$$

22. Evaluate the coefficients in the series

$$R = \int Q\left[a^2\frac{\partial\Omega}{\partial Z}\right]dE$$

(constants of integration are determined later). If, from step 21, we write Q in the form

$$Q = \sum_{k,l} Q_{k,l}\sin(kE+lH)$$

and if $a^2(\partial\Omega/\partial Z)$ is written as

$$a^2\frac{\partial\Omega}{\partial Z} = \sum_{i,j}\left[C_{i,j}\cos(iE-j\phi) + S_{i,j}\sin(iE-j\phi)\right] ,$$

then R may be written as

$$R = c_0 E + c_1 E \cos H + s_1 E \sin H$$

$$+ \sum_{i,j} \Big[A_{-1,i,j} \cos(iE - j\phi - H) + B_{-1,i,j} \sin(iE - j\phi - H)$$

$$+ A_{0,i,j} \cos(iE - j\phi) + B_{0,i,j} \sin(iE - j\phi)$$

$$+ A_{+1,i,j} \cos(iE - j\phi + H) + B_{+1,i,j} \sin(iE - j\phi + H) \Big] ;$$

where

$$c_0 = \frac{1}{2} \sum_k \left(S_{k,0} - S_{-k,0} \right) Q_{k,0} ,$$

$$c_1 = \frac{1}{2} \sum_k \left(S_{k,0} - S_{-k,0} \right) Q_{k,1} ,$$

$$s_1 = \frac{1}{2} \sum_k \left(C_{k,0} + C_{-k,0} \right) Q_{k,1} ,$$

$$A_{-1,i,j} = \frac{1}{2\left(i - j\,\frac{n'}{n}\right)} \sum_k C_{i+k,j}\, Q_{k,1} ,$$

$$B_{-1,i,j} = \frac{1}{2\left(i - j\,\frac{n'}{n}\right)} \sum_k S_{i+k,j}\, Q_{k,1} ,$$

$$A_{0,i,j} = \frac{1}{2\left(i - j\,\frac{n'}{n}\right)} \sum_k \left(C_{i+k,j} - C_{i-k,j} \right) Q_{k,0} ,$$

$$B_{0,i,j} = \frac{1}{2\left(i - j\,\frac{n'}{n}\right)} \sum_k \left(S_{i+k,j} - S_{i-k,j} \right) Q_{k,0} ,$$

$$A_{+1,i,j} = \frac{1}{2\left(i - j\,\frac{n'}{n}\right)} \sum_k - C_{i-k,j}\, Q_{k,1} ,$$

$$B_{+1,i,j} = \frac{1}{2\left(i - j\,\frac{n'}{n}\right)} \sum_k - S_{i-k,j}\, Q_{k,1}$$

except that for $i=j=0$ all these coefficients are zero. For $j=0$ the coefficients with negative values of i are combined with the coefficients with positive values of i. That is, $A_{-1,-i,0}$ is added to $A_{+1,i,0}$ etc., where i is positive.

23. Evaluate the coefficients in the series

$$u = \bar{R} ;$$

The bar operator means that u is obtained from R by considering the argument H to be the same as E and adding corresponding coefficients. Thus, u will contain

$$c_0 E + c_1 E \cos E + s_1 E \sin E$$

plus pure periodic terms where, for example, the coefficient of $\cos(iE-j\phi)$ will be

$$A_{-1,i+1,j} + A_{0,i,j} + A_{+1,i-1,j} .$$

24. Evaluate the coefficients in the expressions

$$\frac{M}{3} = -\frac{1-\frac{1}{2}e^2}{(1-e^2)} + \frac{2e}{3(1-e^2)}\cos(E) - \frac{e^2}{6(1-e^2)}\cos(2E)$$

$$- \frac{e}{(1-e^2)}\cos(H) + \frac{e^2}{3(1-e^2)}\cos(H+E) + \frac{4-e^2}{3(1-e^2)}\cos(H-E)$$

$$- \frac{e}{3(1-e^2)}\cos(H-2E)$$

and

$$N = \frac{e}{(1-e^2)}\sin(E) - \frac{e^2}{2(1-e^2)}\sin(2E) + \frac{e^2}{(1-e^2)}\sin(H+E)$$

$$- \frac{e}{(1-e^2)}\sin(H) - \frac{2-e^2}{(1-e^2)}\sin(H-E) + \frac{e}{(1-e^2)}\sin(H-2E) .$$

25. Evaluate the coefficients in the series

$$W = \int \left[\frac{M}{3} \cdot \frac{\partial(3a\Omega)}{\partial E} + N \cdot ar \frac{\partial \Omega}{\partial r}\right] dE .$$

If

$$\frac{M}{3} = \sum_{k,l} M_{k,l} \cos(kE + lH) ,$$

$$N = \sum_{k,l} N_{k,l} \sin(kE + lH) ,$$

$$\frac{\partial(3a\Omega)}{\partial E} = \sum_{i,j} \left[C_{i,j} \cos(iE - j\phi) + S_{i,j} \sin(iE - j\phi) \right] ,$$

and

$$ar \frac{\partial \Omega}{\partial r} = \sum_{i,j} \left[c_{i,j} \cos(iE - j\phi) + s_{i,j} \sin(iE - j\phi) \right] ,$$

then W may be written in the same form as **R** in step 22 with

$$c_0 = \frac{1}{2} \sum_k \left[\left(C_{k,0} + C_{-k,0} \right) M_{k,0} + \left(s_{k,0} - s_{-k,0} \right) N_{k,0} \right] ,$$

$$c_1 = \frac{1}{2} \sum_k \left[\left(C_{k,0} + C_{-k,0} \right) M_{k,1} + \left(s_{k,0} - s_{-k,0} \right) N_{k,1} \right] ,$$

$$s_1 = \frac{1}{2} \sum_k \left[\left(- S_{k,0} + S_{-k,0} \right) M_{k,1} + \left(c_{k,0} + c_{-k,0} \right) N_{k,1} \right] ,$$

$$A_{-1,i,j} = \frac{1}{2\left(i - j\frac{n'}{n}\right)} \sum_k \left(- S_{i+k,j} M_{k,1} + c_{i+k,j} N_{k,1} \right) ,$$

$$B_{-1,i,j} = \frac{1}{2\left(i - j\frac{n'}{n}\right)} \sum_k \left(C_{i+k,j} M_{k,1} + s_{i+k,j} N_{k,1} \right) ,$$

$$A_{0,i,j} = \frac{1}{2\left(i - j\frac{n'}{n}\right)} \sum_k \left[\left(- S_{i+k,j} - S_{i-k,j} \right) M_{k,0} + \left(c_{i+k,j} - c_{i-k,j} \right) N_{k,0} \right] ,$$

$$B_{0,i,j} = \frac{1}{2\left(i - j\frac{n'}{n}\right)} \sum_k \left[\left(C_{i+k,j} + C_{i-k,j} \right) M_{k,0} + \left(s_{i+k,j} - s_{i-k,j} \right) N_{k,0} \right] ,$$

$$A_{+1,i,j} = \frac{1}{2\left(i - j\frac{n'}{n}\right)} \sum_k \left(- S_{i-k,j} M_{k,1} - c_{i-k,j} N_{k,1} \right) ,$$

$$B_{+1,i,j} = \frac{1}{2\left(i - j\frac{n'}{n}\right)} \sum_k \left(C_{i-k,j} M_{k,1} - s_{i-k,j} N_{k,1} \right) .$$

26. Evaluate the coefficients in $\bar{W}$ from W by considering H to be the same as E and combining coefficients as in step 23.

27. Evaluate the coefficients in the series

$$\nu = -\frac{(X_0 + eX_1)}{6} - \frac{\bar{W}}{2} ,$$

where X_0 is that part of the W series which does not contain H in the argument and X_1 is the remaining part of the W series with the temporary argument H set equal to zero. Thus, for example, the coefficient of $\cos(iE-j\phi)$ in ν would be

$$-\frac{1}{6}\left[A_{0,i,j} + e\left(A_{-1,i,j} + A_{+1,i,j}\right)\right] - \frac{1}{2}\left[A_{-1,i+1,j} + A_{0,i,j} + A_{+1,i-1,j}\right] .$$

28. Evaluate the coefficients in the series

$$n\delta z = \int \bar{W}(1 - e\cos E)\, dE .$$

The coefficient of E^2 in $n\delta z$ should be zero.

29. Evaluate the coefficients in the series for $d\nu/dE$ and du/dE from the corresponding series for ν and u by differentiating term by term.
30. Convert the terms factored by E in each series to terms factored by time by replacing E by $nt+e\sin E$.
31. Compute numerical values of $n\delta z$, ν, u, $\bar{W}$, $d\nu/dE$, and du/dE corresponding to the epoch of osculation by evaluating the series with E set equal to E_0. Denote these values by zero subscripts.
32. Evaluate the constants of integration:

$$l_1 = -u_0 \frac{\cos E_0}{1 - e\cos E_0} + \left(\frac{du}{dE}\right)_0 \frac{\sin E_0}{1 - e\cos E_0}$$

$$l_2 = -u_0 \frac{\sin E_0}{1 - e\cos E_0} - \left(\frac{du}{dE}\right)_0 \frac{\cos E_0 - e}{1 - e\cos E_0}$$

$$k_1 = -\left[4\bar{W}_0 + 6\nu_0\right] \frac{\cos E_0}{1 - e\cos E_0} - 2\left(\frac{d\nu}{dE}\right)_0 \frac{\sin E_0}{1 - e\cos E_0}$$

$$k_2 = -\left[4\bar{W} + 6\nu_0\right] \frac{\sin E_0}{1 - e\cos E_0} + 2\left(\frac{d\nu}{dE}\right)_0 \frac{\cos E_0 - e}{1 - e\cos E_0}$$

$$k_0 = -k_1\cos E_0 - k_2\sin E_0 - \bar{W}_0$$

and

$$C - g_0 = -\left[\left(1 - \frac{1}{2}e^2\right)\sin E_0 - \frac{e}{4}\sin 2E_0\right] k_1 + \left[\cos E_0 - \frac{e}{4}\cos 2E_0\right] k_2$$

$$- (n\delta z)_0 .$$

33. Add the constants of integration to the series coefficients:

a. In $n\delta z$ add

$$(C - g_0) + \left(k_0 - \frac{1}{2} ek_1\right) nt + \left[\left(1 - \frac{e^2}{2}\right) k_1\right] \sin E$$

$$+ \left(-\frac{e}{4} k_1\right) \sin 2E + (-k_2) \cos E + \left(\frac{e}{4} k_2\right) \cos 2E ;$$

b. In ν add

$$\left(-\frac{2}{3} k_0 - \frac{e}{6} k_1\right) + \left(-\frac{1}{2} k_1\right) \cos E + \left(-\frac{1}{2} k_2\right) \sin E ;$$

c. In u add

$$(-l_1 e) + (l_1) \cos E + (l_2) \sin E .$$

34. Print out the coefficients in the series for the perturbations.

Appendix B

SOME FIRST ORDER PERTURBATIONS

First order general perturbations due to Jupiter are given for six minor planets:

(13)	Egeria
(1286)	Banachiewizca
(132)	Aethra
(241)	Germania
(1274)	Delportia
(1373)	1935 QN.

The orbital elements of the minor planets were taken from the Ephemerides of Minor Planets for 1962.[B1] The elements of Jupiter were taken from Clemence[B2] using the values given for 1950:

$M' = 302^\circ.36489 + 0^\circ.08308578116$ (JD–2433282.0)

$\omega' = 274.14275$

$\Omega' = 99.80204$ } Ecliptic and Mean Equinox 1950.0

$i' = 1.30710$

$e' = 0.04846063$

$a' = 5.20298098$ a.u.

$n' = 0^\circ.08308578116$/day

$m' = 1/1047.355$

The value of m' is taken from the American Ephemeris and Nautical Almanac.

(13) Egeria

Epoch 1938 Dec 8.0 ET = JD 2429240.5
T = 0.0001 (JD–2429240.5)

$M_0 = 31°.864$

$\omega = 78.013$
$\Omega = 43.563$ Ecliptic and Mean Equinox 1950.0
$i = 16.537$

$e = 0.086199424$
$a = 2.5770$ a.u.
$n = 0°.23825639$ per day

$M_0' = 326.57371$

A	B	C
−1.2775339	−2.1720939	+0.5054949
+1.6594002	−1.3084250	−1.4705535
+1.5017412	−0.4020153	+2.0549637

Egeria is the planet for which Hansen originally computed general perturbations[B3] and later published tables.[B4] Further investigations and comparisons with observations were made by Hoelling[B5] and Samter.[B6–B7]

Perturbations of Egeria

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
0	0	5363 T	0 T	8 T	0 T	−19 T	
1	0	−293 T	−105 T	92 T	−256 T	224 T	−1557 T
2	0	6 T	2 T				
0	0	363	0	−218	−0	22	0
1	0	−715	631	−545	−622	−106	−403
2	0	17	−20	9	2	−2	−1
−2	1	−1	1	1	1	0	0
−1	1	17	−11	−11	−19	14	11
0	1	−27	46	8	−8	28	52
1	1	−321	112	−68	−196	−24	−37
2	1	8	1	−3	2	−5	−24
3	1	−2	−0	−0	−3	−0	1
−1	2	1	−1	0	−3	1	−0
0	2	36	33	30	−31	−23	43
1	2	−462	−492	217	−190	58	−61
2	2	−379	−523	546	−396	17	−25
3	2	10	11	2	1	−5	3
4	2	0	1	−1	0	0	−0
−1	3	1	1	−0	−0	0	−0
0	3	−20	−50	−52	21	18	−33
1	3	−1860	−1658	−237	106	59	16
2	3	26	2029	−1728	20	−457	−196
3	3	−37	7	−63	−46	−4	−1
4	3	0	−1	−0	−1	1	1
0	4	−0	−1	−1	0	−0	−1
1	4	2	−7	−6	−0	6	2
2	4	−10	90	−56	−13	−14	−20
3	4	−48	15	−20	−53	1	−16
4	4	13	4	−6	17	0	1
1	5	2	−1	−2	−1	2	−0
2	5	7	53	−15	−2	−2	−2
3	5	−39	2	−4	−38	12	−15
4	5	7	8	−10	8	−3	1
5	5	−0	−4	5	0	0	−0
1	6	−1	1	1	1	−2	0
2	6	0	55	7	3	1	−2
3	6	41	−8	6	34	−14	10
4	6	1	5	−5	2	−3	−1
5	6	2	−3	3	2	1	1
6	6	−1	0	−0	−2	−0	0

Perturbations of Egeria—Continued

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
3	7	5	−1	0	3	−1	1
4	7	1	2	−3	1	−1	−1
5	7	1	−1	1	1	−0	1
6	7	−1	−0	0	−1	0	−0
3	8	5	−1	0	1	−0	0
4	8	1	2	−2	1	−1	−1
5	8	0	−0	0	1	−0	0
6	8	−0	−0	0	−1	0	−0
7	8	0	0	−1	0	−0	−0
3	9	1	−0	0	−0	−0	−0
4	9	−1	−2	1	−1	0	0

(1286) Banachiewizca

Epoch 1951 Dec 20.0 ET = JD 2434000.5
T = 0.0001 (JD−2434000.5)

$M_0 = 215°.072$

$\omega = 100.709$
$\Omega = 201.315$ Ecliptic and Mean Equinox 1950.0
$i = 9.707$

$e = 0.093256863$
$a = 3.0219$ a.u.
$n = 0°.18762278$ per day

$M_0' = 2.06202$

A	B	C
+1.5869835	+2.5537900	−0.1852088
−2.5133799	+1.4943933	−0.7496642
−0.5443193	+0.5453495	+2.9215716

The general perturbations of Banachiewizca were previously computed and given by Herget[B8] as an example.

Perturbations of Banachiewizca

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
0	0	−9587 T	0 T	−25 T	0 T	12 T	
1	0	−633 T	309 T	−271 T	−552 T	−127 T	−1733 T
2	0	15 T	−7 T				
0	0	−4814	0	447	−0	10	0
1	0	−3496	−1822	1616	−3058	110	−84
2	0	81	36	9	−1	−3	2
−2	1	1	0	0	−1	−0	−0
−1	1	−16	4	7	17	6	−20
0	1	88	52	31	1	−21	−102
1	1	856	−257	139	457	4	60
2	1	−1	4	1	17	18	47
3	1	2	−1	2	3	−1	−1
−1	2	−2	−1	2	0	−0	−1
0	2	68	83	105	−74	−156	21
1	2	−2884	−5998	1094	−605	92	−49
2	2	−2380	−3346	3172	−2250	144	−37
3	2	58	87	−4	−2	−10	8
4	2	1	1	−1	1	0	−0
0	3	−5	−1	−1	8	6	−6
1	3	−87	−34	−11	76	−11	−40
2	3	853	−103	94	577	90	93
3	3	131	−195	233	174	6	13
4	3	−5	5	2	−1	−4	−3
5	3	−0	1	−1	−1	0	0
0	4	−0	−0	0	0	0	−1
1	4	−0	11	13	5	−15	−9
2	4	361	−372	132	74	21	1
3	4	−162	−358	357	−150	46	−78
4	4	56	27	−24	70	−3	2
5	4	−1	−1	1	1	1	−2
1	5	−1	−1	−2	2	3	−0
2	5	2	−59	−21	11	0	−10
3	5	190	78	−52	146	23	43
4	5	34	−35	37	47	13	3
5	5	−1	13	−25	−0	−1	−1
6	5	0	−0	−1	1	1	−0
2	6	−1	1	1	2	−1	−3
3	6	66	−14	11	24	5	3
4	6	1	−48	50	4	12	−12
5	6	12	9	−11	16	0	4
6	6	−6	2	−3	−9	0	−1

Perturbations of Banachiewizca—Continued

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
2	7	−0	−1	−1	−0	1	1
3	7	56	−66	−12	−4	1	−3
4	7	46	64	−51	38	1	23
5	7	8	−4	2	11	3	2
6	7	−2	5	−7	−2	−1	0
7	7	−1	−2	3	−2	0	0
3	8	−0	−0	−0	0	0	−1
4	8	11	4	−1	6	1	1
5	8	4	−6	6	5	3	−1
6	8	2	3	−4	2	−0	1
7	8	−2	0	0	−3	−0	−0
8	8	1	−1	1	1	0	0
3	9	−1	1	1	1	−1	−3
4	9	1189	−314	14	20	−1	3
5	9	−7	−102	90	−5	16	−31
6	9	2	3	−1	2	1	1
7	9	−1	1	−1	−1	−1	0
8	9	−0	−1	1	−0	0	−0
9	9	0	0	−0	1	−0	0
5	10	1	2	−1	1	0	0
6	10	1	−1	0	1	1	0
7	10	0	1	−1	−0	−0	0
8	10	−1	−0	0	−1	−0	−0
9	10	0	−0	0	1	0	0
5	11	4	1	−0	1	−0	0
6	11	1	−2	2	1	1	−0
6	12	0	1	−0	0	−0	0

(132) *Aethra*

Epoch 1925 Jan 10.0 ET = JD 2424160.5
T = 0.0001 (JD–2424160.5)

$M_0 = 145°.191$

$\omega = 253.349$
$\Omega = 259.662$
$i = 25.161$
Ecliptic and Mean Equinox 1950.0

$e = 0.38276405$
$a = 2.6123$ a.u.
$n = 0°.23344222$ per day

$M_0' = 264.49794$

A	B	C
−2.0941885	−1.0306821	−1.0926239
+1.4719146	−1.8667986	−0.7579083
−0.5214907	−1.1300793	+2.2485237

Aethra is an interesting planet having a large eccentricity and inclination. Its general perturbations were previously computed by Herget.[B9] Accurate special perturbations using Musen's method[B10] were included in a differential correction by Musen[B11] and revealed certain errors in the observations. There is a term of long period corresponding to $i = 5$, $j = 14$.

Perturbations of Aethra

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
0	0	16316 T	0 T	−353 T	0 T	182 T	
1	0	581 T	980 T	−923 T	507 T	−475 T	3504 T
2	0	−56 T	−101 T				
0	0	8984	0	−1230	−0	−156	0
1	0	−1097	2087	−1960	−967	268	572
2	0	112	−230	19	10	2	−1
3	0	−1	3	−2	−1	0	1
−3	1	0	−0	−0	−0	1	−0
−2	1	−1	10	6	1	1	3
−1	1	7	−57	−67	−6	−75	−48
0	1	−3	−59	−59	15	−136	−97
1	1	65	50	−37	64	177	134
2	1	33	−14	11	42	33	34
3	1	−10	3	−3	−8	−2	−1
4	1	1	−1	0	0	0	−1
−2	2	−3	1	0	2	−1	1
−1	2	−10	2	−14	−20	−4	−24
0	2	125	−110	−105	−276	−119	−287
1	2	−2237	282	−175	−999	83	191
2	2	−375	100	−119	−509	40	139
3	2	43	−17	5	−14	−1	−7
4	2	3	1	−1	3	0	−0
−2	3	−1	−0	−0	0	−1	−0
−1	3	52	10	1	−6	4	−4
0	3	−256	10	−72	433	−134	481
1	3	−9949	−4702	−760	1821	−57	198
2	3	2531	1037	−964	2349	231	−781
3	3	−230	−96	−16	37	6	−13
4	3	−1	2	−5	4	−0	1
5	3	−1	−1	1	−1	−0	0
−1	4	2	3	1	−1	1	1
0	4	−4	−3	−16	16	−22	12
1	4	−81	−104	−143	90	−70	58
2	4	260	460	−333	193	86	−70
3	4	23	38	−83	47	29	−19
4	4	−7	−10	3	−3	−2	1
5	4	0	−1	3	−0	−0	−0
0	5	0	6	−4	−0	−4	−2
1	5	−7	−55	−87	−6	−77	−3
2	5	−153	681	−292	−49	29	0
3	5	−23	198	−220	−29	91	1
4	5	4	−35	14	2	−5	−1
5	5	0	1	0	1	0	−0
6	5	−0	1	−1	−1	0	0

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
0	6	3	−3	−0	−1	0	−1
1	6	−11	14	27	25	25	30
2	6	−395	299	123	145	38	25
3	6	228	−213	194	207	−89	−68
4	6	−10	8	15	15	−7	−6
5	6	−3	3	−2	−2	1	1
6	6	−0	−0	−0	−1	−0	0
1	7	0	−0	0	3	−1	3
2	7	−22	3	0	28	3	17
3	7	92	3	1	69	−4	−18
4	7	20	−3	3	28	−3	−13
5	7	−6	1	−1	−4	0	2
6	7	1	−0	0	0	−0	−0
1	8	1	1	−0	0	−1	0
2	8	−10	−6	−12	14	−11	13
3	8	134	156	−46	52	2	1
4	8	53	35	−37	54	13	−30
5	8	−10	−6	2	−4	−1	2
6	8	1	1	−0	1	0	−0
2	9	−0	2	4	−0	4	0
3	9	1	37	21	−2	6	−3
4	9	−7	−39	35	−7	−13	6
5	9	−1	−1	5	−2	−3	1
6	9	1	2	−1	1	1	−0
3	10	−2	3	4	3	3	1
4	10	11	−12	10	7	−2	−1
5	10	2	−5	6	2	−4	−1
6	10	−1	2	−1	−0	1	0
3	11	−3	1	1	5	1	1
4	11	98	1	4	19	2	1
5	11	20	−9	8	20	−7	−9
6	11	−3	2	−1	−1	1	1
4	12	3	2	2	−2	0	−1
5	12	−5	−3	2	−4	−0	2
6	12	−1	0	0	−1	0	1
4	13	0	1	1	−0	0	−0
5	13	−0	−3	2	−1	−0	0
6	13	−1	−1	1	−1	−1	1
4	14	−1	2	3	2	3	2
5	14	252	−281	14	8	2	−1
6	14	5	−19	18	5	−10	1
7	14	−1	2	−0	−0	0	−0
6	15	−1	0	−0	−1	0	0

(241) Germania

Epoch 1951 Dec 20.0 ET = JD 2434000.5
T = 0.0001 (JD–2434000.5)

$M_0 = 237°.440$

$\omega = 74.116$
$\Omega = 271.529$ Ecliptic and Mean Equinox 1950.0
$i = 5.516$

$e = 0.095828379$
$a = 3.0524$ a.u.
$n = 0°.18481972$ per day

$M_0' = 2.06202$

A	B	C
+2.9435085	+0.7494468	−0.2933036
−0.8069138	+2.6685677	−1.2160512
−0.0423471	+1.2444521	+2.7843020

First order general perturbations of Germania were previously computed by Kline and Herget[B12] and compared with special perturbations in order to draw conclusions concerning the accuracy of first order general perturbations.

Perturbations of Germania

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
0	0	18838 T	0 T	−18 T	0 T	−33 T	
1	0	−673 T	220 T	−193 T	−587 T	342 T	−1184 T
2	0	16 T	−5 T				
0	0	−1278	0	−1228	−0	24	0
1	0	−1094	5961	−5205	−960	−136	−77
2	0	28	−147	4	3	−2	0
−1	1	1	9	11	−2	−9	−6
0	1	64	−17	30	−6	−73	−18
1	1	496	−896	476	262	29	16
2	1	4	−6	26	15	27	27
3	1	1	0	1	2	−1	−1
−1	2	−1	2	−0	−1	−0	1
0	2	22	−62	−74	−64	55	82
1	2	−7361	3764	−703	−1251	−58	−49
2	2	−4265	2778	−2607	−4009	−38	−81
3	2	115	−73	3	8	3	10
4	2	0	0	−0	1	−0	−0
0	3	1	1	1	−1	−3	1
1	3	−17	42	35	−8	−2	19
2	3	−309	−316	198	−236	−30	−84
3	3	−277	29	−27	−334	1	−7
4	3	10	−0	−1	3	−1	4
1	4	−2	−2	−3	−1	6	2
2	4	−214	145	−56	−42	−7	−4
3	4	57	233	−226	44	−12	74
4	4	−61	−33	35	−80	2	−1
5	4	2	1	−1	0	−1	1
1	5	0	−0	−0	−0	−0	1
2	5	4	8	2	−6	2	2
3	5	−73	10	−11	−53	−14	−4
4	5	−7	27	−30	−13	−6	8
5	5	−13	−17	24	−19	1	−0
6	5	0	1	−0	−0	−1	0
3	6	−0	19	−7	2	−1	1
4	6	19	7	−9	19	5	5
5	6	−7	5	−6	−9	−3	1
6	6	−1	−8	12	−2	0	0
3	7	5	1	−1	−1	0	−0
4	7	−6	12	−10	−4	−3	2
5	7	2	2	−3	2	0	1
6	7	−3	1	−0	−4	−1	−0
7	7	1	−3	5	2	0	0

Perturbations of Germania—Continued

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
4	8	2	1	−0	1	0	0
5	8	2	−1	1	2	1	−0
6	8	0	1	−1	0	−0	1
7	8	−1	−0	1	−2	−0	−0
8	8	1	−1	2	2	−0	0
4	9	7	−7	−1	−0	−0	−0
5	9	2	5	−4	2	0	2
6	9	0	−0	−0	0	0	0
7	9	−0	0	−1	−0	−0	0
8	9	−0	−0	1	−0	−0	−0
9	9	1	−0	0	1	−0	0
5	11	−0	−1	0	−0	0	0
6	11	−1	−0	0	−0	−0	−0

(1274) Delportia

Epoch 1932 Oct 20.0 ET = JD 2427000.5
T = 0.0001 (JD–2427000.5)

$M_0 = 214^\circ.74119$

$\omega = 242.98033$
$\Omega = 327.42878$ } Ecliptic and Mean Equinox 1950.0
$i = 4.40990$

$e = 0.1130022$
$a = 2.2290079$ a.u.
$n = 0^\circ.29616788$ per day

$M_0' = 140.46156$

A	B	C
−1.9192056	+1.1226357	−0.0922681
−0.9698111	−1.7192724	−1.0167650
−0.5870198	−0.8299533	+1.9814519

First order general perturbations of Delportia were previously computed by Herget[B13] and used to produce the ephemeris presently in use.

Perturbations of Delportia

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
0	0	−6491 T	0 T	−5 T	0 T	13 T	
1	0	−648 T	48 T	−42 T	−565 T	−113 T	431 T
2	0	18 T	−1 T				
0	0	95	0	106	−0	−8	0
1	0	570	256	−214	497	36	−36
2	0	−16	−8	1	0	1	0
−1	1	−1	2	3	1	−7	3
0	1	12	79	3	2	−11	6
1	1	71	251	−166	47	14	−3
2	1	−2	−4	−2	0	5	−0
−1	2	−0	0	0	0	−1	1
0	2	0	8	−9	8	−8	8
1	2	204	341	−186	113	15	−16
2	2	135	209	−235	152	5	−2
3	2	−5	−7	0	0	−1	0
0	3	−1	2	−4	4	−5	6
1	3	284	388	−102	80	3	−3
2	3	248	280	−269	237	13	−24
3	3	−27	−26	23	−26	−1	0
4	3	1	1	0	−0	0	−0
0	4	0	−0	1	−1	1	−2
1	4	74	81	24	−24	1	−3
2	4	−157	−134	110	−129	−6	18
3	4	−14	−8	14	−22	−1	2
4	4	5	2	−3	7	0	−0
1	5	1	1	1	−2	0	−1
2	5	−23	−16	10	−16	−0	2
3	5	−12	−5	6	−13	−0	2
4	5	5	1	−2	6	0	−0
5	5	−1	−0	0	−2	−0	0
2	6	−10	−5	2	−5	−0	0
3	6	−9	−3	3	−9	0	1
4	6	3	0	−0	3	−0	−0
5	6	−1	−0	−0	−2	0	0
6	6	0	−0	0	1	0	−0
2	7	−102	−39	2	−7	−0	−0
3	7	−29	−5	4	−26	1	4
4	7	2	0	−0	1	−0	−0
5	7	−1	0	−0	−1	0	0
6	7	0	−0	0	1	−0	−0
3	8	2	0	−0	1	−0	−0
4	8	1	−0	0	1	−0	−0

(1373) 1935 QN

Epoch 1941 Jan 6.0 ET = JD 2430000.5
$T = 0.0001$ (JD–2430000.5)

$M_0 = 293°.612$

$\omega = 99.051$
$\Omega = 298.068$ Ecliptic and Mean Equinox 1950.0
$i = 38.902$

$e = 0.32158820$
$a = 3.4111$ a.u.
$n = 0°.15644444$ per day

$M_0' = 29.71890$

A	B	C
+2.0607314	−1.8497219	−1.8902002
+0.7243470	+2.5384346	−1.9809114
+2.6199827	+0.7530860	+2.0343883

(1373) was chosen as an example because it is one of the most interesting minor planets. It has a large eccentricity and inclination and is the only known planet for which there exists a libration in the argument of perihelion.[B14] The secular perturbations have been computed by Smith[B15] using Halphen's method. Smith's results verify the libration in perigee and disclose large secular perturbations in the eccentricity and inclination. There is no dominant small divisor for small values of the indices, and the series for the perturbations converges very slowly. The term $i = 8$, $j = 15$ will contain some inaccuracy since the series are not computed beyond $j = 15$. There will be another significant term at $i = 9$, $j = 17$. A further study of this planet will be given in the future.

Perturbations of (1373)

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
0	0	10279 T	0 T	292 T	0 T	80 T	
1	0	3220 T	−987 T	908 T	2810 T	−249 T	−3505 T
2	0	−259 T	84 T				
0	0	40369	0	−925	−0	183	0
1	0	−2563	1577	−1496	−2220	−334	−1959
2	0	197	−178	64	−14	16	−9
3	0	−2	2	3	−5	−7	3
4	0	1	2	−3	2	0	1
5	0	0	−0	−0	1	1	−0
−4	1	0	0	−0	−1	1	0
−3	1	1	−4	−6	−1	1	−1
−2	1	−6	−3	4	6	−17	−0
−1	1	20	71	111	−24	14	23
0	1	69	292	115	0	−191	85
1	1	342	545	−284	208	149	−78
2	1	100	120	−154	125	75	−35
3	1	−24	−33	27	−19	3	−6
4	1	−2	0	2	−5	−2	2
5	1	1	1	−1	1	−0	1
6	1	0	−0	−0	1	0	−0
−4	2	−0	0	1	0	−0	0
−3	2	1	0	0	−1	1	0
−2	2	−0	−4	−7	1	3	−3
−1	2	−107	−127	−6	14	−19	1
0	2	761	1108	1385	−1073	−941	998
i	2	31997	28101	3784	−3847	−88	47
2	2	−3985	−4468	4120	−3720	668	−673
3	2	402	416	−38	66	6	−4
4	2	−12	−8	8	−10	1	−4
5	2	−2	0	1	−4	−1	1
6	2	1	0	−0	1	−0	1
−3	3	0	0	1	−0	−0	0
−2	3	1	1	1	−2	2	0
−1	3	−2	−6	−8	4	4	−7
0	3	5	15	4	27	−38	9
1	3	−161	−20	−22	229	77	−218
2	3	1494	335	−212	786	−91	208
3	3	137	115	−146	245	−14	30
4	3	5	−8	−6	35	1	2
5	3	−5	−2	2	−4	0	−2
6	3	−1	0	−0	−3	−0	0

Perturbations of (1373)—Continued

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
−2	4	0	0	1	−0	−0	1
−1	4	2	0	0	−3	2	0
0	4	−0	−12	−8	7	4	−12
1	4	−50	49	78	83	−145	33
2	4	511	2041	368	290	44	−210
3	4	1460	−98	115	1206	−26	413
4	4	−85	24	−15	40	−1	−0
5	4	8	−3	3	16	1	2
6	4	−2	−0	1	−1	−0	−1
7	4	−1	0	−1	−1	−0	−0
−1	5	0	0	1	−0	−0	1
0	5	3	−0	−0	−3	2	0
1	5	−11	1	−7	13	0	−21
2	5	−16	−26	−47	31	52	2
3	5	121	394	−185	−23	−39	−26
4	5	−165	50	−73	−168	−14	−55
5	5	19	−2	−4	8	−2	−1
6	5	4	−3	4	6	1	1
7	5	−0	−0	1	−0	−0	−0
8	5	−0	0	−1	−1	−0	−0
0	6	0	0	1	−0	−0	1
1	6	6	−2	−2	−4	3	0
2	6	−31	4	−1	43	−15	−65
3	6	−568	158	−14	155	45	29
4	6	21	247	−198	21	−83	−56
5	6	−25	−1	−23	−27	−4	−8
6	6	3	1	−3	2	−1	−1
7	6	1	−2	2	2	0	0
8	6	−0	−0	1	0	0	−0
1	7	0	1	1	−0	0	1
2	7	−1	−2	−4	−4	6	0
3	7	17	−17	−25	−30	14	36
4	7	−319	186	−37	−110	−7	−24
5	7	−1	−68	60	−16	32	12
6	7	−3	11	−8	−6	−1	−2
7	7	1	1	−2	1	−1	−0
8	7	0	−1	1	1	0	0
2	8	1	1	1	−1	0	2
3	8	1	−9	−14	−4	17	−2
4	8	−16	−119	−54	−13	0	14
5	8	−90	62	−44	−68	6	−20
6	8	3	−15	11	−5	6	1
7	8	−1	3	−3	−1	−0	−1
8	8	0	1	−1	1	−0	0

Perturbations of (1873)—Continued

I	J	$(n\delta z \cdot 10^4)$ degrees		$\nu \cdot 10^6$		$\mu \cdot 10^6$	
		cos	sin	cos	sin	cos	sin
2	9	−0	0	0	0	−1	0
3	9	1	−1	1	−1	1	2
4	9	6	9	15	−9	−19	4
5	9	−79	−205	62	−11	14	−2
6	9	29	−15	24	26	−5	12
7	9	−4	−1	3	−3	2	−0
8	9	−0	1	−1	−0	0	−0
3	10	−0	0	0	0	−1	0
4	10	2	−1	1	−3	2	4
5	10	26	−8	4	−15	−8	1
6	10	−23	−43	30	−15	10	0
7	10	5	0	4	3	−1	2
8	10	−1	−1	1	−1	1	−0
4	11	1	0	1	0	−1	0
5	11	−3	3	5	6	−3	−8
6	11	116	−72	6	28	2	5
7	11	14	23	−20	17	−9	−0
8	11	−0	−3	1	1	−0	1
9	11	−0	−0	0	−1	0	−0
5	12	0	1	1	−0	−2	1
6	12	3	8	5	0	−0	−3
7	12	14	−9	5	10	−0	4
8	12	1	2	−2	2	−1	−0
6	13	−2	−2	−4	4	5	−4
7	13	104	118	−18	8	−3	3
8	13	−16	9	−11	−13	1	−7
9	13	2	−1	−0	1	−0	−0
6	14	−0	0	−0	0	−0	−1
7	14	−2	1	−0	2	1	−1
8	14	5	6	−4	3	−1	0
9	14	−1	1	−1	−1	−0	−1
7	15	2	−2	−2	−4	1	9
8	15	−374	242	−5	−19	−3	−5
9	15	−10	−18	15	−11	8	6
10	15	1	2	−0	−0	0	−0

REFERENCES

B1. Ephemerides of Minor Planets for 1962.

B2. CLEMENCE, G. M., "On the Elements of Jupiter," *Astronom. J.*, **52**(1160): 89–93, 1946.

B3. HANSEN, P. A., "Auseinandersetzung einer zweck-Mässigen Methode zur Berechnung der Absoluten Störungen der Kleiner Planeten," Leipzig, 1857–59.

B4. ———, "Tafeln der Egeria," *Abhandlungen der Mathematisch-Physischen Classe der Kgl. Sachsischen Gesellschaft der Wissenschaften in Leipzig*, **8**, 393–569, 1867.

B5. HOELLING, J. H., "Untersuchungen über die Bewegung des Planeten (13) Egeria," *Astronomische Abhandlungen* (A.N. Nr 12), 11–30, 1907.

B6. SAMTER, H., Uber die Bahn des Planeten Egeria (13), *Ak. d. Wiss*, 1239–1253, 1909.

B7. ———, Die Bewegung des Planeten (13) Egeria, *Astronomische Abhandlungen* (A.N. Nr. 3), 9–30.

B8. HERGET, P. H., The Computation of Orbits, 1948.

B9. ———, The Orbit and Perturbations of (132) Aethra, *Astronom. J.*, **47**, 1938.

B10. MUSEN, P., Special Perturbations in the Vectorial Elements, *Astronom. J.*, **59**, 1954.

B11. ———, New Elements of (132) Aethra, Minor Planet Circular 1577, 1957 IV4.

B12. KLINE, J. E., and HERGET, P., "On the Accuracy of First Order General Perturbations," *Astronom. J.*, **49**, 121, 1940.

B13. HERGET, P., Elements and General Perturbations of (1274) Delportia, *Astronom. J.*, **47**, 122, 1938.

B14. KOZAI, Y., Secular Perturbations of Asteroids with High Inclination and Eccentricity, *Astronom. J.*, **67**, 1962.

B15. SMITH, A. J., Jr., "Halphen's Method of Secular Perturbations Applied to the Determination of Long Range Effects in the Motion of Celestial Bodies, Part II," NASA Technical Report (to be published).

A DEVELOPMENT OF THE EARTH'S GRAVITATIONAL POTENTIAL FROM THE SIXTH THROUGH THE TWELFTH ZONAL HARMONIC

THEODORE L. FELSENTREGER

Goddard Space Flight Center

INTRODUCTION

The purpose of this report is to express that part of the earth's gravitational potential from the sixth zonal harmonic through the twelfth, along with its partial derivatives with respect to an inertial geocentric coordinate system, in a form suitable for inclusion in the numerical integration of position partial derivatives for an earth satellite. Values for the harmonic coefficients are also given.

EQUATIONS OF MOTION

The earth's potential is

$$U=(\mu/r)\left[1-\sum_{n=2}^{\infty}(R/r)^n J_n P_n(\sin\phi)\right],$$

where

$\mu=GM$
$R=$ radius of the earth
$J_n=$ zonal harmonic coefficients $(n=2, 3, \cdots)$
$P_n=$ Legendre polynomials $(n=2, 3, \cdots)$
$\phi=$ geocentic latitride.

The Legendre polynomials are

$$P_6(\sin\phi)=\tfrac{1}{16}(231\sin^6\phi-315\sin^4\phi+105\sin^2\phi-5)$$

$$P_7(\sin\phi)=\tfrac{1}{16}(429\sin^7\phi-693\sin^5\phi+315\sin^3\phi-35\sin\phi)$$

$$P_8(\sin\phi)=\tfrac{1}{128}(6435\sin^8\phi-12012\sin^6\phi+6930\sin^4\phi-1260\sin^2\phi+35)$$

$$P_9(\sin\phi)=\tfrac{1}{128}(12155\sin^9\phi-25740\sin^7\phi+18018\sin^5\phi-4620\sin^3\phi+315\sin\phi)$$

$$P_{10}(\sin\phi)=\tfrac{1}{256}(46189\sin^{10}\phi-109395\sin^8\phi+90090\sin^6\phi-30030\sin^4\phi+3465\sin^2\phi-63)$$

$$P_{11}(\sin\phi)=\tfrac{1}{256}(88179\sin^{11}\phi-230945\sin^9\phi+218790\sin^7\phi-90090\sin^5\phi+15015\sin^3\phi-693\sin\phi)$$

$$P_{12}(\sin\phi)=\tfrac{1}{1024}(676039\sin^{12}\phi-1939938\sin^{10}\phi+2078505\sin^8\phi-1021020\sin^6\phi$$
$$+225225\sin^4\phi-18018\sin^2\phi+231).$$

We consider an orthogonal, earth-centered inertial coordinate system in which the z-axis coincides with the earth's axis of rotation. Then,

$$z = r \sin\phi$$

$$r = (x^2+y^2+z^2)^{1/2}.$$

For the disturbing function

$$F = (\mu/r) \sum_{n=6}^{12} (R/r)^n J_n P_n(\sin\phi),$$

we have

$$\begin{aligned}
F = (\mu/r)[&(J_6R^6/16r^{12})(231z^6-315z^4r^2+105z^2r^4-5r^6)\\
&+(J_7R^7z/16r^{14})(429z^6-693z^4r^2+315z^2r^4-35r^6)\\
&+(J_8R^8/128r^{16})(6435z^8-12012z^6r^2+6930z^4r^4-1260z^2r^6+35r^8)\\
&+(J_9R^9z/128r^{18})(12155z^8-25740z^6r^2+18018z^4r^4-4620z^2r^6+315r^8)\\
&+(J_{10}R^{10}/256r^{20})(46189z^{10}-109395z^8r^2+90090z^6r^4-30030z^4r^6+3465z^2r^8-63r^{10})\\
&+(J_{11}R^{11}z/256r^{22})(88179z^{10}-230945z^8r^2+218790z^6r^4-90090z^4r^6+15015z^2r^8-693r^{10})\\
&+(J_{12}R^{12}/1024r^{24})(676039z^{12}-1939938z^{10}r^2+2078505z^8r^4-1021020z^6r^6\\
&+225225z^4r^8-18018z^2r^{10}+231r^{12})],
\end{aligned}$$

$$\begin{aligned}
\partial F/\partial x = -(\mu x/r^3)[&(7J_6R^6/16r^{12})(429z^6-495z^4r^2+135z^2r^4-5r^6)\\
&+(9J_7R^7z/16r^{14})(715z^6-1001z^4r^2+385z^2r^4-35r^6)\\
&+(45J_8R^8/128r^{16})(2431z^8-4004z^6r^2+2002z^4r^4-308z^2r^6+7r^8)\\
&+(55J_9R^9z/128r^{18})(4199z^8-7956z^6r^2+4914z^4r^4-1092z^2r^6+63r^8)\\
&+(33J_{10}R^{10}/256r^{20})(29393z^{10}-62985z^8r^2+46410z^6r^4-13650z^4r^6+1365z^2r^8-21r^{10})\\
&+(39J_{11}R^{11}z/256r^{22})(52003z^{10}-124355z^8r^2+106590z^6r^4-39270z^4r^6+5775z^2r^8-231r^{10})\\
&+(91J_{12}R^{12}/1024r^{24})(185725z^{12}-490314z^{10}r^2-479655z^8r^4-213180z^6r^6\\
&+42075z^4r^8-2970z^2r^{10}+33r^{12})],
\end{aligned}$$

$$\begin{aligned}
\partial F/\partial y = -(\mu y/r^3)[&(7J_6R^6/16r^{12})(429z^6-495z^4r^2+135z^2r^4-5r^6)\\
&+(9J_7R^7z/16r^{14})(715z^6-1001z^4r^2+385z^2r^4-35r^6)\\
&+(45J_8R^8/128r^{16})(2431z^8-4004z^6r^2+2002z^4r^4-308z^2r^6+7r^8)\\
&+(55J_9R^9z/128r^{18})(4199z^8-7956z^6r^2+4914z^4r^4-1092z^2r^6+63r^8)\\
&+(33J_{10}R^{10}/256r^{20})(29393z^{10}-62985z^8r^2+46410z^6r^4-13650z^4r^6+1365z^2r^8-21r^{10})\\
&+(39J_{11}R^{11}z/256r^{22})(52003z^{10}-124355z^8r^2+106590z^6r^4-39270z^4r^6+5775z^2r^8-231r^{10})\\
&+(91J_{12}R^{12}/1024r^{24})(185725z^{12}-490314z^{10}r^2+479655z^8r^4-213180z^6r^6\\
&+42075z^4r^8-2970z^2r^{10}+33r^{12})],
\end{aligned}$$

$$\partial F/\partial z = -(\mu/r^3)\left[(7J_6R^6z/16r^{12})(429z^6-693z^4r^2+315z^2r^4-35r^6)\right.$$
$$+(J_7R^7/16r^{14})(6435z^8-12012z^6r^2+6930z^4r^4-1260z^2r^6+35r^8)$$
$$+9J_8R^8z/128r^{16})(12155z^8-25740z^6r^2+18018z^4r^4-4620z^2r^6+315r^8)$$
$$+(5J_9R^9/128r^{18})(46189z^{10}-109395z^8r^2+90090z^6r^4-30030z^4r^6+3465z^2r^8-63r^{10})$$
$$+(11J_{10}R^{10}z/256r^{20})(88179z^{10}-230945z^8r^2+218790z^6r^4-90090z^4r^6+15015z^2r^8-693r^{10})$$
$$+(3J_{11}R^{11}/256r^{22})(676039z^{12}-1939938z^{10}r^2+2078505z^8r^4-1021020z^6r^6$$
$$+225225z^4r^8-18018z^2r^{10}+231r^{12})$$
$$+(13J_{12}R^{12}z/1024r^{24})(1300075z^{12}-4056234z^{10}r^2+4849845z^8r^4-2771340z^6r^6$$
$$\left.+765765z^4r^8-90090z^2r^{10}+3003r^{12})\right].$$

HARMONIC EFFICIENTS

Values for the zonal harmonic coefficients as given by Kozai[1] are

$$J_6 = 0.39 \pm 0.12 \times 10^{-6}$$
$$J_7 = -0.470 \pm 0.021 \times 10^{-6}$$
$$J_8 = -0.02 \pm 0.02 \times 10^{-6}$$
$$J_9 = 0.117 \pm 0.025 \times 10^{-6}.$$

Values presented by Kind-Hele, Cook, and Rees[2] are

$$J_6 = 0.72 \pm 0.2 \times 10^{-6}$$
$$J_8 = 0.34 \pm 0.2 \times 10^{-6}$$
$$J_{10} = -0.50 \pm 0.2 \times 10^{-6}$$
$$J_{12} = 0.44 \pm 0.2 \times 10^{-6}.$$

(A value for J_{11} could not be found.)

REFERENCES

1. Kozai, Y., "Numerical Results from Orbits," Smithsonian Inst., Astrophys. Observ. Spec. Rept. No. 101, July 31, 1962.
2. King-Hele, D. G., Cook, G. E., and Rees, J. M., "Determination of the Even Harmonics in the Earth's Gravitational Potential," *Geophys. J. of the Royal Astro. Soc.*, **8**(1): 119–145, September 1963.

COMPARISON OF THE VON ZEIPEL AND MODIFIED HANSEN METHODS AS APPLIED TO ARTIFICIAL SATELLITES

DAVID FISHER

Goddard Space Flight Center

The solutions to the problem of the near earth satellite without drag obtained by applying the von Zeipel method and the modified Hansen method are compared. Formulas are derived for osculating elements when the modified Hansen theory is expressed in terms of orbital true longitude. Differences in the arbitrary constants are tabulated. Transformations that relate the time element of the two theories are also given.

INTRODUCTION

Widely different theories are often used in computing orbits of artificial satellites. It is of interest to examine the results of different theories when they are applied to the basic problem of the near earth satellite without drag. Of special importance are the major theories of celestial mechanics introduced by Brouwer[1] and by Musen[2,3] in solving this problem.

Brouwer[1] applied the method of von Zeipel to the near earth satellite problem and obtained analytic representations for the osculating Delaunay and Keplerian elements. The results are given by Brouwer to order J_2 in the elements and J_2^2 in the mean motions, where J_2 is the coefficient of the second zonal harmonic of the earth's potential, and equals the quantity $2k_2$ appearing in the articles of Brouwer and Musen. Musen,[2,3] on the other hand, first modified Hansen's method, then by applying it to the same problem of the near searth atellite without drag, showed how to obtain the psoition of the satellite in a semi-analytic manner to any prescribed order of J_2. The solution of the satellite problem in terms of orbital true longitude by Musen[3] is considered below.

The results obtained by Brouwer are given in a form convenient for comparison with the results of many authors. Indeed Kozai,[4] Garfinkel[5] and others have readily compared their solutions with Brouwer's. However, since Musen's formulations of the problem are intended to provide numerical results of high precision for the position of a satellite, explicit analytic formulations of the perturbations of the elements do not appear in his articles. For that reason, formulas are given here for elements derived from the modified Hansen theory in terms of orbital true longitude so that the results of Musen and Brouwer can be compared.

As would be expected, the differences between the two theories are exhibited in the respective choices of the arbitrary constants and in the arguments of the trigonometric terms. The constants of both theories are discussed and presented in tabular form. The transformations of the variables of the angular arguments are presented. Therefore, when the solutions to the satellite problem are carried out to the same order in J_2 by the methods of Brouwer and Musen, full correspondence can be obtained by taking into account the differences in the constants and the angular variables.

THE OSCULATING ELEMENTS

The definitions of the osculating elements appearing in Brouwer's article may be found in

any text on celestial mechanics—for example, Brouwer and Clemence.[6] It is, moreover, a relatively simple matter to find expressions for the osculating elements of the modified Hansen theory as expressed in terms of orbital true longitude; these differ from the corresponding formulas of the modified Hansen theory as expressed in terms of eccentric anomaly given by Bailie and Bryant,[7] since the W functions differ slightly. To indicate how such representations of osculating elements are derived, we shall now review briefly some of the concepts and definitions of the modified Hansen theory as expressed in terms of orbital true longitude.

DEFINITIONS FROM THE MODIFIED HANSEN THEORY

When the differential equations given in Musen's articles[2,3] are solved, expressions for the components Ξ, Υ, and Ψ, of the $\bar{W}$ function, for the λ parameters, and for the perturbation of the pseudo-time $n_0\,\delta z$ result. The functions Ξ, Υ, and Ψ are expressed in terms of orbital true longitude and are related to osculating elements by the formulas

$$\left.\begin{aligned} \Xi &= -1 - \frac{h_0}{h} + 2\frac{h}{h_0}\,, \\ \Upsilon &= 2\frac{h}{h_0}\,e\cos\phi - \left(1 + \frac{h_0}{h}\right)e_0\,, \\ \Psi &= 2\frac{h}{h_0}\,e\sin\phi\,. \end{aligned}\right\} \tag{1}$$

Here $-\phi$ is the deviation of the osculating true anomaly from the true anomaly of the auxiliary ellipse, e the osculating eccentricity, and h is proportional to the reciprocal of the Delaunay variable G; that is,

$$G=\mu/h. \tag{2}$$

The quantities h_0 and e_0 are the elements of Hansen's auxiliary ellipse.

The λ parameters are defined by the formulas

$$\left.\begin{aligned} \lambda_1 &= \sin\tfrac{i}{2}\cos N & \lambda_3 &= \cos\tfrac{i}{2}\sin K \\ \lambda_2 &= \sin\tfrac{i}{2}\sin N & \lambda_4 &= \cos\tfrac{i}{2}\cos K. \end{aligned}\right\} \tag{3}$$

Here i is the osculating angle of inclination of the orbit plane and corresponds to I in Brouwer's development. The quantities K and N are Fourier series of the order of the perturbations and do not contain secular terms.

The angular variables are given by the formulas

$$\left.\begin{aligned} f &= cv - \pi_0 - \phi \quad , \\ \omega &= (g - c)v + (\pi_0 - \theta_0) + \phi + K + N \quad , \\ \theta &= (1 - h')v + \theta_0 + K - N \quad . \end{aligned}\right\} \tag{4}$$

The quantities f, ω, and θ are the osculating true anomaly, argument of perigee, and longitude of the node, respectively. The quantities g, c, and h' in the right-hand side of equations 4 are proportional to the mean motions of the argument of latitude, mean anomaly, and the longitude of the ascending node, respectively. The quantities π_0 and θ_0 are prescribed constants.

The time element of the auxiliary ellipse is denoted by the symbol z and is often called the pseudo-time. When orbital true longitude is the argument, the mean anomaly of the auxiliary ellipse is $c(n_0)_H z$. The symbol n_0 appears with different meanings in the articles of Brouwer and Musen; therefore, the symbol $(n_0)_H$ is adopted here instead of the n_0 appearing in Musen's article. The quantity δz is the deviation of the pseudo-time from the unperturbed satellite time.

OSCULATING ELEMENTS FOR THE MODIFIED HANSEN THEORY

By inverting equations 1 and 3 it is readily found that

$$\left.\begin{aligned} G &= \frac{\mu}{h} = \frac{\mu}{h_0}\left(1 - \frac{\Xi}{3} + \frac{2}{27}\Xi^2 + \ldots\right) \quad , \\ e &= e_0 + \frac{1}{2}(T - e_0\Xi) + \frac{1}{24}\left(4e_0\Xi^2 - 4\Xi T + \frac{3\Psi^2}{e_0}\right) + \ldots \quad , \\ \frac{H}{G} &= \cos i = \cos i_0\left(1 + \frac{\Xi}{3} + \frac{\Xi^2}{27} + \ldots\right) \quad . \end{aligned}\right\} \tag{5}$$

Similarly, the quantities associated with the angular variables are found to be

$$\left.\begin{aligned} \phi &= \frac{\Psi}{2e_0} + \frac{\Psi}{4e_0}\left[\frac{\Xi}{3} - \frac{T}{e_0}\right] + \ldots \quad , \\ K + N &= \frac{\lambda_3}{\cos\frac{i_0}{2}} + \frac{\lambda_2}{\sin\frac{i_0}{2}} - \frac{\Xi}{12}\cos i_0\left[\frac{\lambda_3}{\cos^3\frac{i_0}{2}} - \frac{\lambda_2}{\sin^3\frac{i_0}{2}}\right] + \ldots \quad , \\ K - N &= \frac{\lambda_3}{\cos\frac{i_0}{2}} - \frac{\lambda_2}{\sin\frac{i_0}{2}} - \frac{\Xi}{12}\cos i_0\left[\frac{\lambda_3}{\cos^3\frac{i_0}{2}} + \frac{\lambda_2}{\sin^3\frac{i_0}{2}}\right] + \ldots \quad . \end{aligned}\right\} \tag{6}$$

It is instructive to derive an expression for the perturbation of the radius vector of the satellite using equations 5 and 6. This leads to an important result already given in the modified Hansen theory. If u denotes the reciprocal of the radius vector of the satellite, and δ the deviation of an osculating element from its value in the auxiliary ellipse, then

$$\delta u = u - \bar{u} = h_0 \frac{\partial u}{\partial h}\delta\left(\frac{h}{h_0}\right) + \frac{\partial u}{\partial e}\delta e + \frac{\partial u}{\partial f}\delta f \quad . \tag{7}$$

TABLE 1.—*Constants Appearing in the Satellite Theories (Order J_2)*

Quantity	Brouwer's Notation (von Zeipel method)	Musen's Notation (modified Hansen method)
$G = \frac{\mu}{h}$	G''	$\frac{\mu}{h_0}\left(1 - \frac{c_0}{3}\right)$
e	$e'' - \frac{\mu^2 J_2 (1 - 3\cos^2 I'')}{8 e'' G''^4}\left(5 - 3\eta''^2 - 2\eta''^3\right)$	$e_0 + \frac{c_1 - e_0 c_0}{2}$
$\frac{H}{G} = \cos i$	$\cos I''$	$\cos i_0 \left(1 + \frac{c_0}{3}\right)$
mean motion of mean anomaly	$\frac{dl''}{dt}$	$c\,(n_0)_H$

from Ξ and Υ. Constants also appear in the solution for the elements by Brouwer. In order to compare the two theories the constants of the elements G, e, and cos$i$ with respect to true anomaly are found from Brouwer's development to order J_2. Similarly, constants with respect to orbital true longitude are found in the article of Bailie and Fisher[8] and are listed in table 1.

The constants appearing in Table 1 are defined as follows:

$$\left.\begin{aligned}
\eta'' &= \sqrt{1 - e''^2}\ , \\
\frac{dl''}{dt} &= n_0 - \frac{3}{4}\frac{n_0 \mu^2 J_2}{L' G''^3}\left(1 - 3\cos^2 I''\right)\ , \\
c &= 1 + \frac{3}{4} J_2 \frac{h_0^4}{\mu^2}\left(1 - 3\cos^2 i_0\right)\ , \\
c_0 &= \frac{3}{4} J_2 \frac{h_0^4}{\mu^2}\left(1 - 3\cos^2 i_0\right)\left(4 - 2\sqrt{1 - e_0^2}\right)\ , \\
c_1 &= \frac{1}{4}\frac{J_2}{e_0}\frac{h_0^4}{\mu^2}\left(1 - 3\cos^2 i_0\right)\left[4\left(1 - \sqrt{1 - e_0^2}\right) + 3e_0^2 - 2e_0^2\sqrt{1 - e_0^2}\right]\ .
\end{aligned}\right\} \qquad (11)$$

These values are taken from the article of Brouwer and from the article of Bailie and Fisher.[8]

The relations between the mean motions of the argument of perigee and the longitude of the node in the articles of Brouwer and Musen are given by the formulas

$$\left.\begin{aligned}
\frac{dg''}{dt} &= (n_0)_H\,(g - c)\ , \\
\frac{dh''}{dt} &= (n_0)_H\,(1 - h')\ .
\end{aligned}\right\} \qquad (12)$$

But we have

$$\left.\begin{aligned} u &= \frac{h^2}{\mu}(1 + e\cos f) \ , \\ \frac{\partial u}{\partial h} &= 2\frac{h}{\mu}(1 + e\cos f) \ , \\ \frac{\partial u}{\partial e} &= \frac{h^2}{\mu}\cos f \ , \\ \frac{\partial u}{\partial f} &= -\frac{eh^2}{\mu}\sin f \ , \\ \bar{u} &= \frac{h_0^2}{\mu}(1 + e_0\cos \bar{f}) \ , \end{aligned}\right\} \tag{8}$$

where $f=\bar{f}-\phi$; and from equations 5 and 6 we have, to order J_2,

$$\left.\begin{aligned} \delta\left(\frac{h}{h_0}\right) &= \frac{\Xi}{3} \ , \\ \delta e &= \frac{\Upsilon - e_0\,\Xi}{2} \ , \\ \delta f &= -\phi = -\frac{\Psi}{2e_0} \ . \end{aligned}\right\} \tag{9}$$

Substituting the required quantities from equations 8 into equation 7 we get, to order J_2,

$$\delta u = \frac{1}{2}\frac{h_0^2}{\mu}\bar{W} + \frac{\Xi}{6}\bar{u} \ , \tag{10}$$

where $\bar{W}=\Xi+\Upsilon\cos\bar{f}+\Psi\sin\bar{f}$, which is consistent with the results of the modified Hansen theory.

COMPARISON OF RESULTS TO THE FIRST ORDER IN J_2

By solving the equations given by Musen, first order analytic solutions for the quantities Ξ, Υ, Ψ, and the λ parameters were obtained by Bailie and Fisher.[8] When the analytic expressions for Ξ, Υ, and Ψ are substituted into equations 5 immediate agreement is obtained with the periodic part of the elements G, e, and I obtained in Brouwer's solution. Similarly, agreement for the periodic part of the expressions for the angular variables ω and θ given by equation 4 with the variables g and h can be readily obtained, when the analytic results of Bailie and Fisher are introduced.

It has now been indicated that the periodic part of the solution of the elements of the satellite problem by Brouwer and Musen agree to the first order in J_2. Although differences in the arbitrary constants and arguments of the trigonometric terms do exist, they do not appear in the first order solutions for the trigonometric parts of the elements since they have J_2 as a multiplier. These differences are exhibited in the terms of the second order and are discussed below.

THE ARBITRARY CONSTANTS OF THE THEORIES

Differences of order J_2 appear in the arbitrary constants of the solutions of the satellite problem by Brouwer and by Musen. The quantities c_0 and $c_1\cos f$ in Musen's work are added to the $\bar{W}$ function and consequently the constants c_0 and c_1 are added to Ξ and Υ; These constants thus occur in the solution for those elements derived

Formulas to order J_2^2 for these mean motions are given in the article of Brouwer and the article of Bailie and Fisher. At first sight the terms in J_2^2 seem to disagree. However, by taking the relationships given in table 1 into account, full agreement is obtained to order J_2^2 in the mean motion of the variables as defined in equations 12.

The differences in the constants given in table 1 will also be exhibited in the coefficients of trigonometric terms of order J_2^2 in the elements derived by the methods of Brouwer and of Musen. Additional differences appear in these coefficients and are due to differences in the *arguments* of the trigonometric terms. We shall now describe these.

THE TIME ELEMENTS OF THE THEORIES

In the von Zeipel method adopted by Brouwer the true anomalies f and f' appear; Brouwer then shows how to relate these true anomalies to the true time of the satellite. In the Hansen method modified by Musen the true anomaly of the auxiliary ellipse $\bar{f}$ (or δ as it is denoted by Bailie and Fisher[8]) appears; Musen shows how to relate $\bar{f}$ to the true time. The true anomalies in the two theories differ by trigonometric terms of order J_2; consequently, it is logical to apply Taylor's theorem to find the relation between these two true anomalies.

We recall that the true anomaly is a function of the eccentricity as well as of the mean anomaly, as is shown by the equation of the center.[6] Also, we have

$$\left.\begin{aligned} f_{osc} &= \bar{f} - \phi \ , \\ e_{osc} &= e_0 + \delta e \ , \end{aligned}\right\} \qquad (13)$$

where f_{osc} and e_{osc} are the osculating true anomaly and osculating eccentricity, while ϕ and δe are of order J_2.

The quantity f appearing in Brouwer's article is related to the osculating mean anomaly l by the equation

$$\frac{df}{dl} = \left(1 - e_0^2\right)^{\frac{-3}{2}} \left(1 + e_0 \cos f\right)^2 \ . \qquad (14)$$

We then find by Taylor's theorem that

$$F(\bar{f}) = F(f) + \left(\phi + \frac{\partial f}{\partial e}\,\delta e\right)\frac{\partial F}{\partial \bar{f}} \ , \qquad (15)$$

where

$$\frac{\partial f}{\partial e} = \left(\frac{2 + e_0 \cos f}{1 - e_0^2}\right) \sin f \ . \qquad (16)$$

Equation 15 transforms a function of the true anomaly of the auxiliary ellipse, $\bar{f}$, to a function of the true anomaly f appearing in Brouwer's article. To extend this transformation so that a function of $\bar{f}$ may be expressed in terms of the mean true anomaly f', we simply apply Taylor's Theorem again to obtain

$$F(\bar{f}) = F(f') + \left(\phi + \frac{\partial f}{\partial e}\,\delta e + \frac{df}{dl}\,\Delta l\right)\frac{\partial F}{\partial \bar{f}} \ , \qquad (17)$$

to the first order in J_2. Here f' is the mean true anomaly in the sense given in Brouwer's article and may be evaluated by Kepler's equation for a given instant of time. The perturbation Δl is the deviation of the mean anomaly from its mean value and is given by the formula $\Delta l = -\partial S_1/\partial L'$ in Brouwer's article. It may also be found from the variation equation in terms of orbital true longitude by the methods adopted in the article of Bailie and Fisher.

In particular, if $F(\bar{f}) = \sin\bar{f}$, we have

$$\sin \bar{f} = \sin f' + \left(\phi + \frac{\partial f}{\partial e}\,\delta e + \frac{df}{dl}\,\Delta l\right) \cos \bar{f} \ . \qquad (18)$$

The multiplier of $\cos\bar{f}$ is of order J_2, so that when f' is given $\bar{f}$ may be found by successive approximations.

It is possible to arrive at equation 17 by approaching the transformation from a somewhat different point of view. From the work of E. Brown[9] we have

$$\delta f = \frac{df}{dl}\,\delta l + \frac{\partial f}{\partial e}\,\delta e \ . \qquad (19)$$

Since the symbol δ refers to the deviation of the osculating element from the corresponding element of the auxiliary ellipse, we have

$$\delta l = l - \left[c(n_0)_H\, z + l_0''\right] \ , \qquad (20)$$

where the quantity in brackets is the mean anomaly of the auxiliary ellipse. Since by table 1

$$c(n_0)_H = dl''/dt \qquad (21)$$

and

$$z - t = \delta z \ ,$$

we have, from equation 20,

$$\delta l = \left[l - \left(\frac{dl''}{dt} t + l_0'' \right) \right] - c(n_0)_H \, \delta z \ . \quad (22)$$

When only the short period terms of the mean anomaly are considered, we have

$$\delta l = \Delta l - c(n_0)_H \, \delta z. \quad (23)$$

Or, substituting in equation 19, we find

$$\frac{df}{dl} c(n_0)_H \, \delta z = \phi + \frac{\partial f}{\partial e} \delta e + \frac{df}{dl} \Delta l \ , \quad (24)$$

which equals the multiplier of $\partial F/\partial \bar{f}$ in equation 17. Consequently, equation 17 may be thought of as transforming a function of the pseudo-time z to a function of time t (to order J_2) by the relation

$$F(z) = F(t) + (\partial F/\partial z)\delta z. \quad (25)$$

In order to complete the transformation, the argument of perigee is taken into consideration. From equations 4 it is seen that a term proportional to the equation of the center must be included in the transformations. It then follows that

$$F(\bar{f}, \bar{\omega}) = F(f', g') + \left(\phi + \frac{\partial f}{\partial e} \delta e + \frac{df}{dl} \Delta l \right) \frac{\partial F}{\partial \bar{f}} + (g - c)(f - l) \frac{\partial F}{\partial \bar{\omega}} \ , \quad (26)$$

where

$$\bar{\omega} = (g - c)v + (\pi_0 - \theta_0) \ .$$

If $F(f')$ and $F(f', g')$ in the right-hand sides of equations 17 and 26 represent periodic terms of order J_2, then the subsequent terms in the right-hand sides of these equations will be of order J_2^2. Thus, by means of equations 17 and 26, it is possible to derive comparisons of the periodic terms of the two theories when they are both developed to order J_2^2.

Kozai[10] has extended the work of Brouwer[1] to order J_2^2 in the periodic terms and J_2^3 in the secular terms. Unfortunately, a corresponding extension has not been made of the analytic results of Bailie and Fisher.[8] However, for the purpose of checking the formulas of the present paper the author has obtained the development of only the short period terms of the Delaunay variable G to order J_2^2 in terms of orbital true longitude. This was done so that comparison could be made with the corresponding terms obtained by Kozai using the method of von Zeipel. For this variable, agreement was obtained.

SUMMARY AND CONCLUSIONS

The solutions to the problem of the near earth satellite without drag given by Brouwer[1] and by Musen[2,3] agree when carried out to the same order in J_2. Due allowance must be made for the differences in the constants and in the ways of expressing the time element.

The differences in the arbitrary constants have been tabulated here to the first order in J_2, and transformations have been given relating the true anomaly of the auxiliary ellipse to that of the satellite.

REFERENCES

1. Brouwer, D., "Solution of the Problem of Artificial Satellite Theory Without Drag," *Astronom. J.*, **64**(9): 378–397, 1959.
2. Musen, P., "Application of Hansen's Theory to the Motion of an Artificial Satellite in the Gravitational Field of the Earth," *J. Geophys. Res.*, **64**(12): 2271–2279, December 1959.
3. ———, "The Theory of Artificial Satellites in Terms of the Orbital True Longitude," *J. Geophys. Res.*, **66**(2): 403–409, February 1961.
4. Kozai, Y., "The Motion of a Close Earth Satellite," *Astronom. J.*, **64**(9): 367–377, November 1959.
5. Garfinkel, B., "The Orbit of a Satellite of an Oblate Planet," *Astronom. J.*, **64**(9): 353–367, November 1959.

6. BROUWER, D., and CLEMENCE, G., "Methods of Celestial Mechanics," New York: Academic Press, 1961.

7. BAILIE, A., and BRYANT, R., "Osculating Elements Derived from the Motion of an Artificial Satellite," *Astronom. J.*, **65**(8): 451–456, October 1960.

8. ———, and FISHER, D., "An Analytic Representation of Musen's Theory of Artificial Satellites in Terms of the Orbital True Longitude," NASA Technical Note D–1468, January 1963.

9. BROWN, E., "An Introductory Treatise on the Lunar Theory," Oxford: Cambridge University Press, 1896.

10. KOZAI, Y., "Second-Order Solution of Artificial Satellite Theory Without Air Drag," *Astronom. J.*, **67**(7): 446–461, September 1962.

Appendix A

ANALYTIC RESULTS FROM THE MODIFIED HANSEN THEORY

By solving the equations given by Musen,[1] first order analytic solutions for the quantities Ξ, Υ, Ψ, λ_2, λ_3, g, $1-h'$, and $g-c$ were obtained by Bailie and Fisher.[2] These are listed here for the convenience of the reader.

To conform to modern notation, the quantities k_2, k_3, and k^8 appearing in the modified Hansen theory are here designated by $J_2/2$, $J_3/2$, and $-J_4/8$, respectively. In formulas A1 through A5 the quantities g and c introduced by integration have been restored, along with the quantity μ; these were set equal to unity by Bailie and Fisher who were then concerned only with terms of order J_2.

$$\begin{aligned}\Xi = {} & \frac{3}{2} J_2 \frac{h_0^4}{\mu^2} (1 - 3\cos^2 i_0)\left[2 - \sqrt{1 - e_0^2}\right] \\ & - \frac{9}{4} J_2 \frac{h_0^4}{\mu^2} (1 - \cos^2 i_0)\left\{\frac{\cos 2\eta}{g} + e_0 \frac{\cos(\xi + 2\eta)}{c + 2g} + e_0 \frac{\cos(\xi - 2\eta)}{2g - c}\right. \\ & \left. - \frac{1}{12}\left[e_0^2 - \frac{10 e_0^2 \cos^2 i_0}{1 - 5\cos^2 i_0}\right]\cos(2\xi - 2\eta)\right\} - \frac{3}{2}\frac{J_3}{J_2}\frac{h_0^2}{\mu} e_0 \sin i_0 \sin(\xi - \eta) \\ & + \frac{15}{16}\frac{J_4}{J_2}\frac{h_0^4}{\mu^2} e_0^2 \left[1 - 3\cos^2 i_0 - \frac{8\cos^4 i_0}{1 - 5\cos^2 i_0}\right]\cos(2\xi - 2\eta) \quad ; \end{aligned} \tag{A1}$$

$$\begin{aligned}\Upsilon = {} & \frac{1}{8} J_2 \frac{h_0^4}{\mu^2} (1 - 3\cos^2 i_0)\left[2e_0 + \frac{8e_0 + 4e_0^3}{1 + \sqrt{1 - e_0^2}}\right] - \frac{1}{8}\frac{J_2}{c}\frac{h_0^4}{\mu^2} (1 - 3\cos^2 i_0)\left[(12 + 3e_0^2)\cos\xi\right. \\ & \left. + 6e_0 \cos 2\xi + e_0^2 \cos 3\xi\right] + \frac{3}{16} J_2 \frac{h_0^4}{\mu^2} (1 - \cos^2 i_0)\left\{8e_0 \frac{\cos 2\eta}{g}\right. \\ & + (28 + 5e_0^2)\frac{\cos(\xi + 2\eta)}{c + 2g} + (4 - e_0^2)\frac{\cos(\xi - 2\eta)}{2g - c} + 12e_0 \frac{\cos(2\xi + 2\eta)}{c + g} \\ & + \frac{1}{3}\left[20e_0 + e_0^3 - \frac{(20e_0 + 10e_0^3)\cos^2 i_0}{1 - 5\cos^2 i_0}\right]\cos(2\xi - 2\eta) + 5e_0^2 \frac{\cos(3\xi + 2\eta)}{3c + 2g} \\ & \left. + e_0^2 \frac{\cos(3\xi - 2\eta)}{3c - 2g}\right\} - \frac{1}{2}\frac{J_3}{J_2}\frac{h_0^2}{\mu}(2 + e_0^2) \sin i_0 \sin(\xi - \eta)\end{aligned}$$

$$+\frac{5}{16}\frac{J_4}{J_2}\frac{h_0^4}{\mu^2}\left(2e_0+e_0^3\right)\left[1-3\cos^2 i_0-\frac{8\cos^4 i_0}{1-5\cos^2 i_0}\right]\cos(2\xi-2\eta)\ ; \tag{A2}$$

$$\begin{aligned}\Psi = &-\frac{1}{8}\frac{J_2}{c}\frac{h_0^4}{\mu^2}\left(1-3\cos^2 i_0\right)\left[\left(12+9e_0^2\right)\sin\xi+6e_0\sin 2\xi+e_0^2\sin 3\xi\right]\\
&+\frac{3}{16}J_2\frac{h_0^4}{\mu^2}\left(1-\cos^2 i_0\right)\Bigg\{12e_0\frac{\sin 2\eta}{g}+\left(28+11e_0^2\right)\frac{\sin(\xi+2\eta)}{c+2g}\\
&+\left(4-7e_0^2\right)\frac{\sin(\xi-2\eta)}{2g-c}+12e_0\frac{\sin(2\xi+2\eta)}{c+g}\\
&+\frac{1}{3}\left[20e_0+e_0^3-20e_0\cos^2 i_0-\frac{\left(100e_0-20e_0^3\right)\cos^4 i_0}{1-5\cos^2 i_0}+\frac{100e_0^3\cos^6 i_0}{\left(1-5\cos^2 i_0\right)^2}\right]\sin(2\xi-2\eta)\\
&+5e_0^2\frac{\sin(3\xi+2\eta)}{3c+2g}+e_0^2\frac{\sin(3\xi-2\eta)}{3c-2g}\Bigg\}\\
&+\frac{J_3}{J_2}\frac{h_0^2}{\mu}\sin i_0\left[1-\frac{2e_0^2\cos^2 i_0}{1-5\cos^2 i_0}\right]\cos(\xi-\eta)\\
&+\frac{5}{16}\frac{J_4}{J_2}\frac{h_0^4}{\mu^2}\Bigg[2e_0+e_0^3-\left(6e_0+5e_0^3\right)\cos^2 i_0\\
&-\frac{\left(16e_0+12e_0^3\right)\cos^4 i_0}{1-5\cos^2 i_0}+\frac{16e_0^3\cos^6 i_0}{\left(1-5\cos^2 i_0\right)^2}\Bigg]\sin(2\xi-2\eta)\ ;\end{aligned} \tag{A3}$$

$$\begin{aligned}\lambda_2 = &\frac{3}{8}J_2\frac{h_0^4}{\mu^2}\sin i_0\cos i_0\cos\frac{i_0}{2}\Bigg\{2e_0\frac{\sin\xi}{c}-\frac{\sin 2\eta}{g}-e_0\frac{\sin(\xi+2\eta)}{c+2g}\\
&+e_0\frac{\sin(\xi-2\eta)}{2g-c}+\Bigg[-\frac{e_0^2\left(8+3\cos i_0-18\cos^2 i_0\right)}{6\left(1-5\cos^2 i_0\right)}\\
&+\frac{e_0^2\left(5+\cos i_0-6\cos^2 i_0\right)\left(1-15\cos^2 i_0\right)}{12\left(1-5\cos^2 i_0\right)^2}\Bigg]\sin(2\xi-2\eta)\Bigg\}\\
&-\frac{1}{4}\frac{J_3}{J_2}\frac{h_0^2}{\mu}e_0\cos i_0\cos\frac{i_0}{2}\left(\frac{1-2\cos i_0-3\cos^2 i_0}{1-5\cos^2 i_0}\right)\cos(\xi-\eta)\end{aligned}$$

$$- \frac{5}{16} \frac{J_4}{J_2} \frac{h_0^4}{\mu^2} e_0^2 \sin i_0 \cos i_0 \cos \frac{i_0}{2} \left[\frac{4 - 7 \cos^2 i_0}{1 - 5 \cos^2 i_0} \right.$$

$$\left. - \frac{(5 + \cos i_0 - 6 \cos^2 i_0)(1 - 7 \cos^2 i_0)}{2(1 - 5 \cos^2 i_0)^2} \right] \sin (2\xi - 2\eta) \quad ; \tag{A4}$$

$$\lambda_3 = - \frac{3}{8} J_2 \frac{h_0^4}{\mu^2} \sin i_0 \cos i_0 \sin \frac{i_0}{2} \left\{ 2e_0 \frac{\sin \xi}{c} - \frac{\sin 2\eta}{g} - e_0 \frac{\sin (\xi + 2\eta)}{c + 2g} \right.$$

$$+ e_0 \frac{\sin (\xi - 2\eta)}{2g - c} + \left[- \frac{e_0^2 (8 - 3 \cos i_0 - 18 \cos^2 i_0)}{6(1 - 5 \cos^2 i_0)} \right.$$

$$\left. \left. + \frac{e_0^2 (5 - \cos i_0 - 6 \cos^2 i_0)(1 - 15 \cos^2 i_0)}{12(1 - 5 \cos^2 i_0)^2} \right] \sin (2\xi - 2\eta) \right\}$$

$$+ \frac{1}{4} \frac{J_3}{J_2} \frac{h_0^2}{\mu} e_0 \cos i_0 \sin \frac{i_0}{2} \left(\frac{1 + 2 \cos i_0 - 3 \cos^2 i_0}{1 - 5 \cos^2 i_0} \right) \cos (\xi - \eta)$$

$$+ \frac{5}{16} \frac{J_4}{J_2} \frac{h_0^4}{\mu^2} e_0^2 \sin i_0 \cos i_0 \sin \frac{i_0}{2} \left[\frac{4 - 7 \cos^2 i_0}{1 - 5 \cos^2 i_0} \right.$$

$$\left. - \frac{(5 - \cos i_0 - 6 \cos^2 i_0)(1 - 7 \cos^2 i_0)}{2(1 - 5 \cos^2 i_0)^2} \right] \sin (2\xi - 2\eta) \quad ; \tag{A5}$$

$$g = 1 + \frac{3}{2} J_2 \frac{h_0^4}{\mu^2} \cos^2 i_0 + \frac{15}{32} J_4 \frac{h_0^8}{\mu^4} (2 + 3e_0^2)(3 \cos^2 i_0 - 7 \cos^4 i_0)$$

$$+ \frac{3}{32} J_2^2 \frac{h_0^8}{\mu^4} \left[\left(110 - 48 \sqrt{1 - e_0^2} + 13e_0^2\right) \cos^2 i_0 - \left(290 - 144 \sqrt{1 - e_0^2} + 9e_0^2\right) \cos^4 i_0 \right] ; \tag{A6}$$

$$1 - h' = - \frac{3}{2} J_2 \frac{h_0^4}{\mu^2} \cos i_0 - \frac{15}{32} J_4 \frac{h_0^8}{\mu^4} (2 + 3e_0^2)(3 \cos i_0 - 7 \cos^3 i_0)$$

$$- \frac{3}{32} J_2^2 \frac{h_0^8}{\mu^4} \left[\left(88 - 40 \sqrt{1 - e_0^2} + 9e_0^2\right) \cos i_0 - \left(236 - 120 \sqrt{1 - e_0^2} + 5e_0^2\right) \cos^3 i_0 \right] ; \tag{A7}$$

$$g - c = -\frac{3}{4} J_2 \frac{h_0^4}{\mu^2} \left(1 - 5\cos^2 i_0\right) - \frac{15}{128} J_4 \frac{h_0^8}{\mu^4} \left[\left(12 + 9e_0^2\right) - \left(144 + 126e_0^2\right) \cos^2 i_0 \right.$$

$$\left. + \left(196 + 189e_0^2\right) \cos^4 i_0 \right]$$

$$- \frac{3}{128} J_2^2 \frac{h_0^8}{\mu^4} \left[\left(162 - 64\sqrt{1 - e_0^2} + 25e_0^2\right) - \left(1500 - 672\sqrt{1 - e_0^2} + 126e_0^2\right) \cos^2 i_0 \right.$$

$$\left. + \left(2810 - 1440\sqrt{1 - e_0^2} + 45e_0^2\right) \cos^4 i_0 \right] \qquad \text{(A8)}$$

where

$$\xi = cv - \pi_0 \text{ and } \eta = gv - \theta_0$$

REFERENCES

1. MUSEN, P., "Application of Hansen's Theory to the Motion of an Artificial Satellite in the Gravitational Field of the Earth," *J. Geophys. Res.*, **64**(12): 2271–2279, December 1959.
2. BAILIE, A., and FISHER, D., "An Analytic Representation of Musen's Theory of Artificial Satellites in Terms of the Orbital True Longitude," NASA Technical Note D–1468, January 1963.

A HYPOTHETICAL FOUR-BODY PROBLEM AND ITS APPLICATIONS: AN ESTIMATE OF THE EFFECT OF THE MOON AND THE SUN ON THE SYNCOM ORBIT

SU-SHU HUANG

Goddard Space Flight Center

A hypothetical system has been constructed by assuming that the center of mass of two finite bodies which are revolving around each other in circular orbit is itself circling with a third finite body at a large distance around the center of mass of the entire system. This somewhat artificial system is an idealization of the earth-moon-sun system. Thus, any satellite that moves in the latter system may be treated, as an approximation, as an infinitesimal fourth body in the hypothetical system, thereby constituting a hypothetical four-body problem.

An investigation of the energy integral of the hypothetical four-body problem confirms that the motion of an artificial satellite in the earth-moon-sun system may be treated with a good approximation, as two three-body problems.

An application of the problem thus formulated is found in the study of the general behavior of the syncom orbit in the presence of the moon as well as the sun. It appears that the presence of both destroys the perfect synchronization of the axial rotation of the earth and the revolution of the syncom satellite around it. It appears that the period of the fourth body around the earth fluctuates with the relative motion of the sun with respect to the earth-moon system. However, the deviation from the perfect synchronization is exceedingly small as would be expected and has no practical significance.

I. FORMULATION OF A HYPOTHETICAL FOUR-BODY PROBLEM

In a previous paper (Huang, 1961) it was suggested that the motion of an artificial satellite in the sun-earth-moon system may be treated as a combination of two restricted three-body problems. The hypothetical four body problem presented in this paper is a mathematical formulation of this suggestion.

Let the masses of the artificial satellite (i.e., the fourth body), the earth, the moon, and the sun be denoted respectively by m, m_1, m_2, and m_3. Since the separation of the earth (m_1) and the moon (m_2) is much smaller than the separation of the earth-moon system as a whole from the sun, and since

$$m_3 \gg m_1+m_2 \gg m \sim 0, \tag{1}$$

we may, when dealing with the general behavior of the artificial satellite, approximate the sun-earth-moon dynamical system by a somewhat artificial system of two masses m_1 and m_2 rotating in circular orbits around their center of mass O' which in turn is rotating in a circular orbit around the center of mass O of the entire system. We denote A_1 and A_2, respectively, the distance from O to m_3 and to O' and write

$$A = A_1 + A_2 \tag{2}$$

Let us consider two rectangular coordinate systems (ξ, η, ζ) and (ξ', η', ζ'); both have their origin at the center of mass O of the three finite bodies and both are fixed in space. The $O\xi'\eta'$ plane represents the orbital plane of O' (which is the center of mass of m_1 and m_2) and m_3 around O while the $O\xi\eta$ plane is parallel to the orbital plane of m_1 and m_2 around O'. We now choose the intersection of these two planes as the ξ-axis

in the (ξ, η, ζ) system and also as the ξ'-axis in the (ξ', η', ζ') system. Consequently the other coordinate axes of both systems are fixed by the right-hand rule.

We now choose m_1+m_2 as the unit of mass, the separation between m_1 and m_2 as the unit of length. Thus, if we denote $m_2=\mu$, we have $m_1=1-\mu$. Also, the unit of time is so chosen that the angular velocity of m_1 and m_2 around O' be unity. It is easy to see that expressed in this system of units, the gravitational constant in Newton's law of gravitation is equal to one.

Thus, the equations of motion of the infinitesimal body m in the (ξ, η, ζ) system are given by

$$\frac{d^2\xi}{dt^2}=-(1-\mu)\frac{\xi-\xi_1}{r_1^3}-\mu\frac{\xi-\xi_2}{r_2^3}-m_3\frac{\xi-\xi_3}{r_3^3} \tag{3}$$

$$\frac{d^2\eta}{dt^2}=-(1-\mu)\frac{\eta-\eta_1}{r_1^3}-\mu\frac{\eta-\eta_2}{r_2^3}-m_3\frac{\eta-\eta_3}{r_3^3} \tag{4}$$

$$\frac{d^2\zeta}{dt^2}=-(1-\mu)\frac{\zeta-\zeta_1}{r_1^3}-\mu\frac{\zeta-\zeta_2}{r_2^3}-m_3\frac{\zeta-\zeta_3}{r_3^3} \tag{5}$$

where r_1, r_2, and r_3 are respectively the distance of m from m_1, m_2, and m_3 while (ξ_1, η_1, ζ_1), (ξ_2, η_2, ζ_2), and (ξ_3, η_3, ζ_3) are respectively coordinates of m_1, m_2, and m_3.

Assuming that ω is the angular velocity of O' and m_3 around O in the $O\xi'\eta'$ plane and α is the angle $O'O\xi'$ at time $t=0$. In the adopted unit system

$$\omega=[(m_1+m_2+m_3)^{1/2}/A^{3/2}] \tag{6}$$

Therefore, the coordinates of O' in the (ξ', η', ζ') can readily be written. By a simple transformation of coordinates we have found that the coordinates of O' in the (ξ, η, ζ) system are:

$$\begin{aligned}\xi_{O'}&=A_2\cos(\omega t+\alpha)\\ \eta_{O'}&=A_2\sin(\omega t+\alpha)\cos i\\ \zeta_{O'}&=A_2\sin(\omega t+\alpha)\sin i\end{aligned} \tag{7}$$

where i is the inclination of planes $O\xi\eta$ and $O\xi'\eta'$. The coordinates (ξ_3, η_3, ζ_3) of m_3 can be obtained simply by replacing A_2 in (7) by $-A_1$. Both α and i are assumed not to change with time.

Since m_1 and m_2 revolve with unit angular velocity around O', their respective coordinates (ξ_1, η_1, ζ_1) and (ξ_2, η_2, ζ_2) in the (ξ, η, ζ) system can now be readily found.

$$\begin{aligned}\xi_1&=\xi_{O'}-\mu\cos t,\\ \eta_1&=\eta_{O'}-\mu\sin t,\\ \zeta_1&=\zeta_{O'},\end{aligned} \tag{8}$$

and

$$\begin{aligned}\xi_2&=\xi_{O'}+(1-\mu)\cos t,\\ \eta_2&=\eta_{O'}+(1-\mu)\sin t,\\ \zeta_2&=\zeta_{O'},\end{aligned} \tag{9}$$

if we start to count time $(t=0)$ when x-axis points to the same direction as the ξ-axis.

Now we furthermore introduce a rotating coordinate system (x, y, z) centered on O'. Let the $O'xy$ plane coincide with the orbital plane of m_1 and m_2 around each other. Hence it is parallel to plane $O\xi\eta$ and the z-axis parallel to the ζ-axis. This system of coordinates is assumed to be rotating with m_1 and m_2. Therefore, we can fix the x-axis such that it will always pass these two finite bodies. The transformation from the (ξ, η, ζ) to the (x, y, z) system is governed by the following equations:

$$\begin{aligned}\xi&=\xi_{O'}+x\cos t-y\sin t,\\ \eta&=\eta_{O'}+x\sin t+y\cos t,\\ \zeta&=\zeta_{O'}+z.\end{aligned} \tag{10}$$

Substituting (9)–(10) into (3)–(5) and utilizing (7)–(9) we can derive the equations of motion of the fourth body in the (x, y, z) system as follows:

$$\frac{d^2x}{dt^2}-2\frac{dy}{dt}=x-\frac{A_2\omega^2x_3}{A}-(1-\mu)\frac{x-x_1}{r_1^3}-\mu\frac{x-x_2}{r_2^3}-m_3\frac{x-x_3}{r_3^3} \tag{11}$$

$$\frac{d^2y}{dt^2}+2\frac{dx}{dt}=y-\frac{A_2\omega^2y_3}{A}-(1-\mu)\frac{y}{r_1^3}-\mu\frac{y}{r_2^3}-m_3\frac{y-y_3}{r_3^3} \tag{12}$$

$$\frac{d^2z}{dt^2}=-\frac{A_2\omega^2z_3}{A}-(1-\mu)\frac{z}{r_1^3}-\mu\frac{z}{r_2^3}-m_3\frac{z-z_3}{r_3^3} \tag{13}$$

where

$$x_1=-\mu, \qquad x_2=1-\mu, \tag{14}$$

and

$$x_3=-A[\cos(\omega t+\alpha)\cos t+\sin(\omega t+\alpha)\cos i\sin t],$$
$$y_3=-A[\sin(\omega t+\alpha)\cos i\cos t-\cos(\omega t+\alpha)\sin t],$$
$$z_3=-A\sin(\omega t+\alpha)\sin i. \tag{15}$$

Equations (11)–(13) may be also expressed in the form

$$\frac{d^2x}{dt^2}-2\frac{dy}{dt}=\frac{\partial U}{\partial x} \tag{16}$$

$$\frac{d^2y}{dt^2}+2\frac{dx}{dt}=\frac{\partial U}{\partial y} \tag{17}$$

$$d^2z/dt^2=\partial U/\partial z \tag{18}$$

if we define

$$U=\tfrac{1}{2}(x^2+y^2)-(m_3/A^3)(x_3x+y_3y+z_3z)+\frac{1-\mu}{r_1}+\frac{\mu}{r_2}+\frac{m_3}{r_3} \tag{19}$$

From (16)–(18) we obtain

$$2U-\left[\left(\frac{dx}{dt}\right)^2+\left(\frac{dy}{dt}\right)^2+\left(\frac{dz}{dt}\right)^2\right]-2\int_0^t\frac{\partial U}{\partial t}\,dt=C \tag{20}$$

where C is an integration constant. The integral in (20) may be transformed to the following form:

$$\int_0^t\frac{\partial U}{\partial t}\,dt=-\int_0^t\left(\frac{m_3}{A^3}-\frac{m_3}{r_3^3}\right)(\mathbf{k}-\boldsymbol{\omega})\cdot(\mathbf{r}\times\mathbf{A})\,dt \tag{21}$$

where $\mathbf{k}$ is a unit vector in the z-direction and represents the angular velocity vector of the revolution of m_1 and m_2 around O', while $\boldsymbol{\omega}$ is the angular velocity vector of the revolution of O' and m_3 around O. In the (x, y, z) coordinate system, the $\boldsymbol{\omega}$ vector has the following three components—$\omega\sin i\sin t$, $-\omega\sin i\cos t$ and $\omega\cos i$. Both $\mathbf{r}$ and $\mathbf{A}$ 5epresent vectors from O′ to m and to m_3 respectively; their magnitudes are r and A.

If we now compare the energy expression given by (20) in the present four-body problem with the well-known Jacobi integral of the restricted three-body problem, we find that the former differs from the latter only by the following three terms:

$$\Delta_1\equiv-2(m_3/A^3)(x_3x+y_3y+z_3z)+(2m_3/r_3)+2\int_0^t\left(\frac{m_3}{A^3}-\frac{m_3}{r_3^3}\right)[(\mathbf{k}-\boldsymbol{\omega})\cdot(\mathbf{r}\times\mathbf{A})]\,dt. \tag{22}$$

In the neighborhood of the earth-moon system r/A is a small quantity. Therefore we may expand these extra terms into a power series of (r/A). Since

$$r_3^{-1}=A^{-1}\sum_{n=0}(r/A)^n p_n(\cos\Theta), \tag{23}$$

where $p_n(x)$ is the Legendre polynomial and Θ is the angle between two vectors $\mathbf{r}$ and $\mathbf{A}$, it can be easily shown that

$$\Delta_1=2(m_3/A)[1+\tfrac{1}{2}(r/A)^2(3\cos^2\Theta-1)]-6\left(\frac{m_3}{A}\right)\int_0^t\left(\frac{r}{A}\right)^2\cos\Theta\,|\sin\Theta|\,(\mathbf{k}-\boldsymbol{\omega})\cdot\frac{\mathbf{r}\times\mathbf{A}}{|\mathbf{r}\times\mathbf{A}|}\,dt, \tag{24}$$

if we neglect the third and higher order terms of (r/A).

In the case of the earth (m_1)—moon (m_2)—sun (m_3) system, ω has a magnitude of 0.07474. Consequently the scalar product of three vectors that appear in the integrand is of the order of unity and both terms on the right side of (24) are of the same order of magnitude with respect to (r/A).

Equation (24) shows that the system of zero-velocity curves of the restricted three-body problem (e.g. Darwin 1897) is not changed by the presence of the third finite body if we maintain an accuracy to the order of r/A, since $2m_3/A$ is a constant and may be absorbed in C. Thus, in the neighborhood of the earth-moon system the motion of an infinitesimal body follows with a good approximation the same kind of restriction as is imposed by the restricted three-body problem even in the presence of the sun.

If m_1, m_2, and m_3 are moving in the same plane, we have the same equations of motion that are given by (11)–(13). The only simplification occurs in the expressions of x_3, y_3, z_3 and Δ_1 which now become

$$x_3=-A\cos[(\omega-1)t+\alpha]$$
$$y_3=-A\sin[(\omega-1)t+\alpha] \tag{25}$$
$$z_3=0$$

and

$$\Delta_1 = (m_3/A)\Big[2+(r/A)^2(3\cos^2\Theta-1) - 3(1-\omega)\int_0^t\left(\frac{r}{A}\right)\sin 2\Theta\, dt\Big]. \quad (26)$$

They are obtained by substituting $i=0$ in (15) and (24) respectively.

II. THE FOUR-BODY PROBLEM IN A DIFFERENT REFERENCE SYSTEM

While in the previous section we have investigated the effect of the sun on the infinitesimal fourth body in the neighborhood of the earth-moon system, we now plan to examine the general behavior of the motion of the fourth body when it is quite remote from the earth-moon system.

In this section and this section alone we will denote m_1 the sun, m_2 the earth, and m_3 the moon and choose a system of units such that both the total mass $m_1+m_2+m_3$ and the separation between m_1 and the center of mass O' of m_2 and m_3 are unity. In a similar way we choose the unit of time so that the gravitational constant G is one. Thus, the angular velocity of m_1 and O' around O is one while that of m_2 and m_3 around O' will be denoted by ω—a situation just reversed of the previous section.

We now write the equation of motion in the (ξ', η', ζ') system which is defined in section I and like the (ξ, η, ζ) system, is at rest in space. We further denote

$$m_1 = 1-\mu, \qquad m_2=\mu_1, \qquad m_3=\mu_2. \quad (27)$$

Hence because of our choice of units $\mu=\mu_1+\mu_2$.

Let the distance between O' and $m_2(\mu_1)$ be a_1 and between O' and $m_3(\mu_2)$ be a_2 and write

$$a=a_1+a_2 \quad (28)$$

which is a small quantity compared to one in our system of units.

The equations of motion in the rest (ξ', η', ζ') system can now be transformed to a (x', y', z') system rotating with m_1 and O'. The origin of the (x', y', z') system is at O and the $x'y'$ plane coincides with the $(\xi'\eta')$ plane with the x'-axis perpetually in the line joining m_1, O, and O' (the positive x' direction is taken to be from O to O'). Thus, the coordinates of m_1 and O' in the (x', y', z') system are respectively $x_1'=-\mu$, $y_1'=z_1'=0$, $x'_{O'}=1-\mu$, $y'_{O'}=z'_{O'}=0$.

The equations of motion in the (x', y', z') system can be easily found from this transformation from the (ξ', η', ζ') system to the (x', y', z') system. They are

$$\frac{d^2x'}{dt^2}-2\frac{dy'}{dt}=x'-(1-\mu)\frac{x'-x_1'}{r_1^3}-\mu_1\frac{x'-x_2'}{r_2^3} - \mu_2\frac{x'-x_3'}{r_3^3} \quad (29)$$

$$\frac{d^2y'}{dt^2}+2\frac{dx'}{dt}=y'-(1-\mu)\frac{y'}{r_1^3}-\mu_1\frac{y'-y_2'}{r_2^3}-\mu_2\frac{y'-y_3'}{r_3^3} \quad (30)$$

$$\frac{d^2z'}{dt^2}=-(1-\mu)\frac{z'}{r_1^3}-\mu_1\frac{z'-z_2'}{r_2^3}-\mu_2\frac{z'-z_3'}{r_3^3} \quad (31)$$

where (x_2', y_2', z_2') and (x_3', y_3', z_3') are the coordinates of m_2 and m_3 and are functions of time. If we denote α the angle that $O'm_3$ makes with $O\xi$ at $t=0$, these coordinates are given by the following expressions:

$$x_2' = 1-\mu-a_1[\cos(\omega t+\alpha)\cos t + \sin(\omega t+\alpha)\cos i \sin t], \quad (32)$$

$$y_2' = a_1[\cos(\omega t+\alpha)\sin t - \sin(\omega t+\alpha)\cos i \cos t], \quad (33)$$

$$z_2' = a_1 \sin(\omega t+\alpha)\sin i. \quad (34)$$

The expression for x_3', y_3', and z_3' can be obtained from (32)–(34) respectively by replacing a_1 by $-a_2$.

Equations (29)–(31) can be reduced to the form as given by (16)–(18) with x, y, z replaced by x', y', z' respectively. However, U is now given by

$$U=\tfrac{1}{2}(x'^2+y'^2)+\frac{1-\mu}{r_1}+\frac{\mu_1}{r_1}+\frac{\mu_2}{r_3} \quad (35)$$

which is slightly different from the corresponding expression given by (19) in the previous section.

Integrating the set of equations of the form

given by (16)–(18) we derive an equation identical to (20) only with x, y, z replaced by x', y', z'. However, the expression of $\partial U/\partial t$ is different in the present case.

Let $\mathbf{a}_1$ be a vector from O' to $m_2(\mu_1)$ and $\mathbf{a}_2$ from O' to $m_3(\mu_2)$. If Θ is the angle between $\mathbf{a}_2$ and the negative x'-direction, we can express $\partial U/\partial t$ in the present case in the following form:

$$\frac{\partial U}{\partial t}=\mu_2 a_2\left(\frac{1}{r_3^3}-\frac{1}{r_2^3}\right)\cdot\{-(\boldsymbol{\omega}-\mathbf{k})\cdot[\mathbf{r}\times(\mathbf{a}_2/a_2)]+(1-\mu)(d/dt)\cos\Theta\}. \tag{36}$$

where $\boldsymbol{\omega}$ and $\mathbf{k}$ denote respectively angular velocity vectors of m_2 and m_3 around O' as well as m_1 and O' around O. Therefore, $\mathbf{k}$ is unit vector in the z' direction.

Comparing the energy integral derived in the present case with that of the restricted three-body problem, we find the following difference:

$$\Delta_2\equiv\frac{2\mu_1}{r_2}+\frac{2\mu_2}{r_3}-\frac{2(\mu_1+\mu_2)}{r_{O'}}-2\int_0^t\frac{\partial U}{\partial t}\,dt \tag{37}$$

where $r'_{O'}$ is the distance of the fourth body from O'. If we combine $m_2(\mu_1)$ and $m_3(\mu_2)$ as a single body of mass $\mu_1+\mu_2=\mu$ place it at O' the four-body problem reduces to the three-body problem with which we have compared our result.

We can expand $1/r_2$ and $1/r_3$ into a series of Legendre polynomials. Taking a_1 and a_2 as small quantities, we obtain the dominating terms for Δ_2 as follows:

$$\Delta_2=a_1a_2\mu\left\{\frac{3\cos^2\varphi-1}{r_{O'}^3}+6\int_0^t\frac{\cos\varphi}{r_{O'}^4}\left[(\boldsymbol{\omega}-\mathbf{k})\cdot\left(\mathbf{r}\times\frac{\mathbf{a}_2}{a_2}\right)-(1-\mu)\frac{d}{dt}\cos\Theta\right]dt\right\} \tag{38}$$

where φ is the angle between $O'm$ and $O'm_3$.

Since we now consider the motion of m at distances far away from m_2 and m_3, all quantities such as $r_{O'}$ and r have the order of magnitude of unity. This leaves Δ_2 to be a second order (a_1a_2) small quantity. It should be noted, however, that ω which enters in (38) directly through $(\boldsymbol{\omega}-\mathbf{k})$ and indirectly through Θ and φ is much greater than unity. But we should also remember that the unit of time adopted here is very large if it is expressed in the c.g.s. units. Therefore in a time interval that is less than a month or so, the integral is still of the order of unity. Over a longer time, the factor $\cos\varphi$ and $\cos\Theta$ will prevent the integral to rapidly increase with time. It follows that the motion of the fourth body in places far away from the earth-moon system are governed, in an approximation correct to the first order of small quantity a, by the same system of zero-velocity curves that are derived from the restricted three-body problem by replacing the earth and the moon by a fictitious mass equal to the sum of the two at their center of mass. In this way we show that the motion of an infinitesimal body in the sun-earth-moon system can indeed be regarded as two three-body problems correct to the first order approximation as we have suggested before (Huang, 1961). In practical cases, however, the motion of an artificial satellite away from the earth-moon system will be affected by other planets and their satellite systems. Consequently the hypothetical four-body problem loses its realistic significance.

III. APPLICATION: ESTIMATE OF THE EFFECT OF THE SYMCOM ORBIT BY THE PRESENCE OF THE MOON AND THE SUN

In order to numerically estimate the effect of the moon and the sun, let us first compute the syncom orbit under the attraction of the earth alone. Needless to say the orbit must be affected by the nonspherical symmetry of the earth but we shall not consider this effect in this paper as it has been extensively studied in recent years (e.g., Brouwer and Clemence, 1961). Thus the syncom orbit around a spherically symmetric earth alone in space is a circle. In the unit system defined in section I, we have found that the period of the syncom circular orbit is 0.22934. This period corresponds to a radius of $a=0.109588$ and a velocity in rest space of $v=3.002369$.

Let us next consider the effect of the moon alone on the syncom orbit. Now in order to have a syncom period of 0.22934 with respect to fixed stars, the period in the rotating frame of reference of the x, y, z system should be 0.23803 because of the advance of the moon during the revolution of the satellite. In other words, we must look for

an orbit of this period in the x, y, z system of the restricted three-body problem. Such an orbit can be easily derived by successive approximation.

Corresponding to the mass ratio of the moon to the total mass of the earth-moon system, μ is taken to be 0.012149 in our computation. A few periodic orbits which are given in table 1 and which are in the neighborhood of the syncom orbit are first found and their periods determined by numerical integration of equations involved in the restricted three-body problem (Huang and Wade, 1963). The integration was performed with double precision by C. Wade, Jr., on the IBM 7094 digital computer at the Goddard Space Flight Center. From the periods of these preliminary orbits we have found, by interpolation, the initial conditions that would lead to the desired orbit with a period of 0.23803, namely the syncom orbit in the presence of the moon. In making the computation we have used only the two dimensional case, assuming that the moon's orbit and the earth's equator lie in the same plane. While such an assumption does not reflect the reality, it does, nevertheless, give an order of magnitude estimate of the effect of the moon on the syncom orbit.

The initial conditions of the syncom orbit in the three-body problem are:

$$x_0 = -0.1217296, \quad y_0 = 0,$$
$$\dot{x}_0 = 0, \quad \dot{y}_0 = -2.893009. \tag{39}$$

When this orbit is plotted, we find that its difference from the circular one, which we would have in the absence of the moon, is very small. The diameter of the orbit along the x-axis is 0.219171 as compared with $2\times 0.109588 = 0.219176$ of the circular one. Therefore, the presence of the moon only makes the diameter along the x-axis of the syncom orbit slightly shorter.

The initial velocity v_0 in the fixed frame of reference is simply $\dot{y}_0 + x_0$ and is equal to -3.014739 from (39). Referred to the center of the earth the initial velocity at x_0 is $\dot{y}_0 + x_0 - x_1$ which is equal to -3.002590. These values should be compared with $v_0 = -3.002369$ for the syncom orbit if the earth were spinning alone in space.

Next we consider the effect of both the moon and the sun on the syncom orbit. The inclusion of the motion of the sun as well as the moon in our search for the syncom orbit naturally leads to the hypothetical four-body problem formulated in section I.

TABLE 1.—*Initial Conditions for Periodic Orbits (in the Restricted Three-body Problem) Which Have Nearly the Syncom Period* $y_0 = \dot{x}_0 = 0$

x_0	$\dot{y}_0$	p_1	Jacobian constant
−0.1216	−2.894916	0.237590	9.7071672
−0.1217	−2.893445	0.237928	9.6992269
−0.1218	−2.891975	0.238266	9.6913067
−0.1219	−2.890507	0.238604	9.6834005
−0.1220	−2.889042	0.238943	9.6755026
−0.1217296	−2.893009	0.238028	9.6968849

Now, an examination of (11)–(13) shows that the equations of motion of the hypothetical four-body problem differ from the corresponding equations in the restricted three-body problem only by the presence of the extra terms

$$m_3/A^3(x_3x + y_3y + z_3z) + (m_3/r_3)$$
$$= (m_3/A)[1 + \tfrac{1}{2}(r/A)^2(3\cos^2\Theta - 1) + \cdots] \tag{40}$$

that occur in the potential function U.

Since U enters into the equations of motion only through its partial derivatives, the first term in the expansion on the right side of (40) produces no effect on the equations. Thus it becomes evident that if the infinitesimal body is moving inside the earth-moon system, as is the case of the syncom satellite, the effect of the sun is small, because the largest relevant term in the perturbing function is of the order of m_3r^2/A^3. For the syncom satellite we may set $r = a = 0.11$. With $m_3 = 3.291\times 10^5$, $A = 3.892\times 10^2$, m_3r^2/A^3 is less than 7×10^{-5}.

We shall assume that the three bodies—the sun, the earth, and the moon—are revolving around each other in the plane of the earth's equator. Therefore, we have numerically studied only the two dimensional case with $\alpha = 0$. The integration of (11)–(12) was performed by C. Wade, Jr., on the IBM 7094 computer, again with double precision. The accuracy of the integration can be checked by the constancy of C in

TABLE 2.—*The Variation of $p_n - p_0$ of Nearly Periodic Orbits in the Hypothetical Four-body Problem Under the Initial Conditions $x_0 = -0.1217296$, $y_0 = \dot{x}_0 = 0$ and $\dot{y}_0$ Listed in the Table ($p_0 = 0.23802760$)*

n (Number of cycle)	$p_n - p_0$			
	$\dot{y}_0 = -2.893009$	$\dot{y}_0 = -2.89298$	$\dot{y}_0 = -2.89302$	$\dot{y}_0 = -2.89304$
1	-0.4455×10^{-5}	-1.1273×10^{-5}	-0.1872×10^{-5}	0.2827×10^{-5}
2	−0.4139	−1.0975	−0.1544	0.3166
3	−0.3561	−1.0435	−0.0961	0.3770
4	−0.2842	−0.9771	−0.0220	0.4571
5	−0.2209	−0.9190	0.0443	0.5264
6	−0.1710	−0.8788	0.0987	0.5860
7	−0.1501	−0.8639	0.1196	0.6143
8	−0.1740	−0.8952	0.1062	0.6024
9	−0.2336	−0.9622	0.0436	0.5472
10	−0.3155	−1.0546	−0.0354	0.4742
11	−0.4109	−1.1589	−0.1367	0.3818
12	−0.5212	−1.2662	−0.2261	0.2805
13	−0.5987	−1.3467	−0.3215	0.2000
14	−0.6434	−1.4033	−0.3632	0.1553
15	−0.6553	−1.4003	−0.3692	0.1434
16	−0.6195	−1.3646	−0.3334	0.1762
17	−0.5540	−1.2811	−0.2738	0.2537
18	−0.4258	−1.1768	−0.1457	0.3431
19	−0.3185	−1.0516	−0.0443	0.4563
20	−0.2351	−0.9384	0.0391	0.5458
21	−0.1457	−0.8490	0.1226	0.6173
22	−0.1159	−0.8192	0.1583	0.6352
23	−0.0980	−0.8013	0.1583	0.6411
24	−0.1457	−0.8430	0.1106	0.5815
25	−0.2231	−0.8967	0.0391	0.5279
26	−0.2947			
27	−0.3722			
28	−0.4318			
29	−0.4735			
30	−0.4496			
31	−0.4318			
32	−0.3662			
33	−0.3066			

(20) in the same way as it is done in the restricted three-body problem by the constancy of the Jacobian constant.

If we now apply the same initial conditions given by (39) that correspond to the syncom orbit in the restricted three-body problem to the present case of the hypothetical four-body problem, we find some interesting facts in our numerical calculation. But before presenting these facts it is necessary to define 'period' for orbits which are not rigorously periodic. We shall define it, in the case of the orbit circling around the earth, as the time interval between two successive crossings over the negative x-axis by the fourth body. Since the period thus defined changes from cycle to cycle for any nearly periodic orbit, we have to distinguish the periods of the different cycles. We shall denote p_n, the nth period, as the interval between $n-1$th and nth crossing of the negative axis.

In table 2 we present the result of our computation, where $(p_n - p_0)$ is given respectively in the

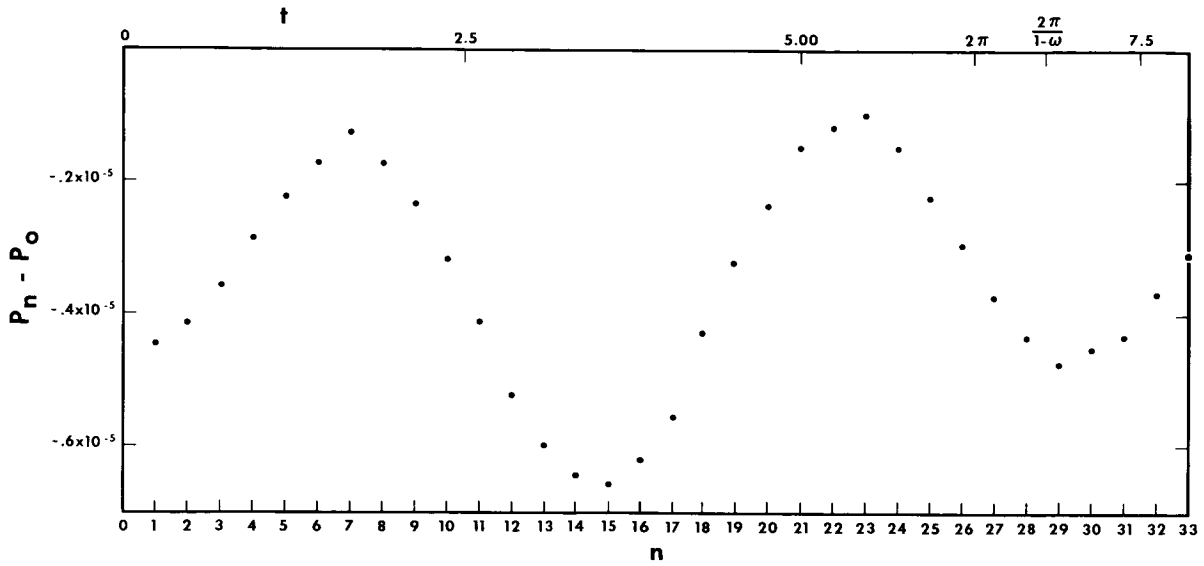


FIGURE 1.—Small fluctuations in period of the syncom orbit due to the presence of both the moon and the sun. In the figure, n denotes the number of cycles that the satellite revolves around the earth since the initial time and is marked on the time axis while $p_n - p_0$ measures the departure of actual periods in different cycles from the perfect synchronization. The period, $2\pi/(1-\omega)$ of the relative motion of the sun with respect to the earth-moon system is also marked on the time axis in order to show that the dominant period of fluctuation is one-half of the periods of the relative solar motion.

second column. p_0 is, of course, the syncom period of 0.2380276 obtained previously in the three-body calculation.

Two immediate conclusions can be drawn from an inspection of table 2. In the first place, the periods are now less than those of the syncom orbits. Secondly, the perturbation by the sun produces small oscillations in the period of the fourth body.

In Fig. 1 we have shown the variation of $p_n - p_0$ with the number of cycle, n, and consequently with time, t. The time axis in the figure denotes time at the end of each cycle. The period $2\pi/(1-\omega)$ of the relative motion of the sun with respect to the earth-moon system is marked on the time axis. Thus, it appears that the variation in $p_n - p_0$ has a dominant period of $\pi/(1-\omega)$ although it is quite evident that it also involves other periods since the amplitude of variation of $p_n - p_0$ is not constant.

We have also integrated the hypothetical four-body problem for a few cases with the same values for x_0, y_0 and $\dot{x}_0$, as given in (39) but with slightly different values for $\dot{y}_0$. The results are given in the third, fourth, and fifth columns of table 2. It appears that the change in $\dot{y}_0$ shifts the variation upward or downward on the $p_n - p_0$ axis as a whole. However, the periodic change is not greatly affected by the small difference in $\dot{y}_0$. An analytical theory for the variation in $p_n - p_0$ is now in progress.

The restricted four-body problem is not limited in its application only to the syncom orbit. Indeed, it may also be used to estimate the effect on the moon's motion due to the perturbation by planets other than the earth, or in general, the effect on motion of any natural satellite by the presence of planets other than its own parent one. Such an approach to the study of motion of a natural satellite follows directly from Hill's (1886; also see Brouwer and Clemence, 1961) original idea that a certain part of the moon's motion that is due to the attraction of the sun may be computed by a simplified model implied in the restricted three-body problem. Similarly the small effect of Jupiter, say, on the motion of the moon may be understood by the hypothetical model implied in the present four-body problem.

Finally, I would like to acknowledge my sincere thanks to Drs. D. Brouwer and G. Hori for their valuable discussions in the course of preparation of this paper. I am also greatly obliged to Mr. C. Wade, Jr., who has performed for me all the numerical integrations involved in this paper on the IBM 7094 computer at Goddard.

REFERENCES

BROUWER, D., and CLEMENCE, G. M., "Method of Celestial Mechanics," New York, Academic Press (1961).

DARWIN, G. H., *Acta Mathematica*, **21,** 99 (1897).

HILL, G. W., *Acta Mathematica*, **8,** 1 (1886).

HUANG, S.-S., *A.J.*, **66,** 157 (1961).

———, and WADE, C., JR., *A.J.*, **68,** 388 (1963).

PERIODIC ORBITS FOR MOON PROBES

SU-SHU HUANG

Goddard Space Flight Center

A general discussion is presented on the broadening of scope, purpose, and—consequently—technique of celestial mechanics as a result of the space age. The discussion is focused on the time scale of the objects being studied. In order to differentiate from the more rigorous part of classical celestial mechanics, the name *Space Mechanics* is suggested to cover this field of study, which is empirical in purpose and numerical in approach.

In the sense of *Space Mechanics*, some periodic orbits that enclose both the two finite bodies have been investigated within the framework of the restricted three-body problem. This represents only the first step in looking for orbits that will permit a moon-probing vehicle to make periodic encounters with the moon on its other side. Two families of periodic orbits have been found—one stable, one unstable—in the orbital plane of the hypothetical moon. Such periodic orbits also have been sought outside the orbital plane. Although a periodic orbit has been obtained numerically, it does not enclose the moon.

TIME SCALES AND METHODS OF APPROACH

Since the coming of the space age, the purpose of celestial mechanics and, consequently, the technique and concepts involved have been drastically broadened. This broadening of the scope results from the time scale of objects under study. Before man introduced artificial satellites and space-probing vehicles, the objects of study in celestial mechanics were confined to celestial bodies that are naturally present in the solar system. These bodies have been in existence, according to most astrophysicists, cosmic chemists, and geophysicists, for a time scale of about 4.5 billion years. With a background of such a long time scale, it would be absurd to talk about an orbit that lasts less than, say, a few thousand years. Indeed, a periodic orbit in celestial mechanics generally was expected to be a mathematically rigorous solution of the equations of motion. Since the problems are so difficult that only men of considerable mathematical talent can make contributions, celestial mechanics naturally becomes a branch of mathematics. In the past centuries many great mathematicians have left their marks in this field.

Our interest in space exploration by means of probing vehicles has modified this situation. We can ask, "What is the time scale of rockets that man on the earth has sent, or will send, out to space?" Without doubt, many of them will last as long as the solar system itself. However, we are not interested in their *entire* life span. For the exploration of space, the upper limit of *useful* rocket life will perhaps be of the order of 100 years. In most cases, the useful life span of a space-probing rocket will be much shorter than this limit.

One hundred years is chosen as a critical time for two reasons. First, it is of the same order of magnitude as the life span of an individual human being. We should remember that the space vehicle is used for experiments whose purpose is to understand the physical nature of an astronomical universe. Whether in physics, chemistry, or biology, an experiment is expected in general to be performed in a length of time shorter than the life span of the investigator who designs the experiment in the earthbound laboratory. This attitude probably will not change drastically for experiments in the spacebound laboratory. Thus a time scale of the order of 100 years for performing an experiment may be regarded as a limit

in most cases of space exploration. Secondly, a round trip in a free orbit from the earth to Pluto, the outermost planet in the solar system, would also be of the order of magnitude of 100 years. Now this does not mean that space travel should be limited forever to the solar system, but the step from interplanetary travel to a visit to our stellar neighbors is wider than the step from Columbus' voyage across the Atlantic Ocean to the astronaut's trip to the moon. Thus considering interstellar travel at the present time is thinking way into the future. For these two reasons, we should be concerned in the next decade or two mainly with orbits of space-probing vehicles having a lifetime of the order of magnitude of 100 years or less.

As a result, the method for attacking the problems in celestial mechanics broadens correspondingly. While it would be most ridiculous to suggest that a numerical solution be attempted for the entire solar system in the next billion years, it is within reason to use the electronic computer for solving many problems connected with rocket trajectories in the solar system during a time interval of a few hundred years or less. Broadening the scope of celestial mechanics as thus understood may be regarded by many investigators, with justification, as its "degeneration." Therefore the name *Space Mechanics* is proposed to cover that part of celestial mechanics used to meet the conditions required by certain kinds of experiments in space research. Thus studies of the trajectories of all kinds of probing vehicles belong to the domain of space mechanics. In this way, *celestial* mechanics will maintain its traditionally high level of mathematical requirements while the numerical results of *space* mechanics will satisfy scientists who design the spacebound experiments.

A PROCEDURE FOR DERIVING PERIODIC ORBITS

With space mechanics in mind, we will talk about periodic orbits for the moon probe. The periodic orbits are supposed to enclose both the earth and the moon and, for practical reasons, to pass around the moon at short distances. Obviously this is a very difficult problem and perhaps has no solution in the rigorous sense. Because of the presence of the sun and the eccentricity of the moon's orbit, it is not certain that we can find such required orbits—which will last for the time interval of a few years. The present paper serves only as a prelininary probing toward this end. It is because we have studied solely the orbits of a test particle with negligible mass in a hypothetical circumstance that the moon is assumed to be revolving in a circular orbit around the earth in the sun's absence. This problem is known as the restricted three-body problem in celestial mechanics (recently reviewed by Szebehely.[1] The periodic orbits thus found[2] encourage us to search for the desired orbits in the actual earth-moon-sun system, although they do not insure that periodic orbits lasting one or more years in the actual system will necessarily be found.

We will undertake our search of periodic orbits in the restricted three-body problem by successive approximation. Thus we first neglect the mass of the moon. Then, if an orbit meets the condition of periodic encounters with the moon, it must satisfy the following equation:

$$p/p_0 = n/m, \tag{1}$$

where p_0 and p are respectively the periods of the moon and the third body and where both n and m are integers. Equation 1 determines the semimajor a of the orbit of the third body; that is,

$$a = (n/m)^{2/3}$$

if the radius of the hypothetical moon's orbit is taken as the unit of length. Because of the requirement that the third body, which will be a moon-probing vehicle, must pass the other side of the moon at a comparatively short distance from it, the semimajor axis a, or equivalently the ratio of integers n/m, must be limited to a certain range of values. Since any two successive close encounters between the moon and the third body take place in a time interval of np_0, a small value of n is preferred in order to have frequent encounters. It is because of these two conditions that there are only a few desirable choices for the value of n/m. The value $n/m = \frac{2}{3}$ proved suitable for our purpose, while Message[3] and Newton[4] have separately studied the case of $n/m = \frac{1}{2}$.

With the ratio n/m (or the semimajor axis a) fixed, we still have a wide choice of possible orbits because of the freedom in assigning values to the orbital eccentricity. Also, the orbital

motion of the third body can be in the same direction (*direct* motion) as, or in the opposite direction (*retrograde* motion) to, the motion of the moon. Therefore we have two families of desired orbits if we consider only close encounters at the apogee of the third body's orbit. In the case where $n/m>1$, another two families of desired orbits may be obtained for encounters at the perigee of the third body's orbit. However encounters at the perigee are not of interest for the present purpose. Thus we shall confine our present study to the two families of orbits resulting from close encounters at the apogee.

The above considerations are based on the assumption that the moon has a negligible mass. We now reason that these two families of periodic orbits may exist even when the mass of the moon, though small, is not negligible. By direct computation two families of periodic orbits have indeed been found for a mass ratio of the two finite bodies corresponding to the earth and the moon.

No analytical proof of the existence of these two families of periodic orbits has been attempted, since the problem is treated as an empirical one and the desired results are derived by numerical experiments. However the procedure by which these orbits are derived may provide some *intuitive* ground to believe in the existence of the periodic solutions in the mathematical sense.

Let us first introduce a rotating coordinate system with the origin at the entire system's center of mass and with the x axis joining the two finite bodies. Let us assume that the third body has the initial conditions given by

$$x=x_0, \quad y=0, \quad \dot{x}=0, \quad \dot{y}=\dot{y}_0.$$

We can then define the period p_n of the nth cycle of a nearly periodic orbit by the time interval between the $n+1$ and nth crossings of the x axis by the third body at about the initial value x_0. Thus p_n can be obtained by interpolation from the results of numerical integration. For a true periodic orbit, it must necessarily be true that

$$p_1=p_2=\cdots p_n=\cdots$$

and the difference between two successive values of p's; that is,

$$\Delta_{n+1,n}=p_{n+1}-p_n$$

measures the deviation from a periodic orbit.

TABLE 1.—*Deviation from Periodicity* ($x_0=-0.39215$)

Direct orbit		Retrograde orbit	
$\dot{y}_0$	Δ_{201}	$\dot{y}_0$	Δ_{201}
−1.6102479	−0.001294	2.3516409	+0.000026
−1.6102480	−0.000457	2.3516410	+0.000009
−1.6102481	+0.000380	2.3516411	−0.000011
−1.6102482	+0.001216	2.3516412	−0.000030

Now we can describe our procedure: First, a value of x_0 is arbitrarily chosen and a few trial values of $\dot{y}_0$ are surmised (from the case when the moon's mass is neglected). We then integrate the equations of motion for the restricted three-body problem and obtain, by interpolation, $\Delta_{2,1}$ for each trial value $\dot{y}_0$. Table 1 illustrates the results for two orbits: one direct, and the other retrograde. We assume the initial condition that yields

$$\Delta_{2,1}=0$$

to be the one that leads to the desired orbit. Thus, the correct initial value of $\dot{y}_0$ should be somewhere between those given in the second and third lines of the table.

A STUDY OF STABILITY

Next we examine the stability of the obtained orbits by investigating the variation in $\Delta_{n+1,n}$

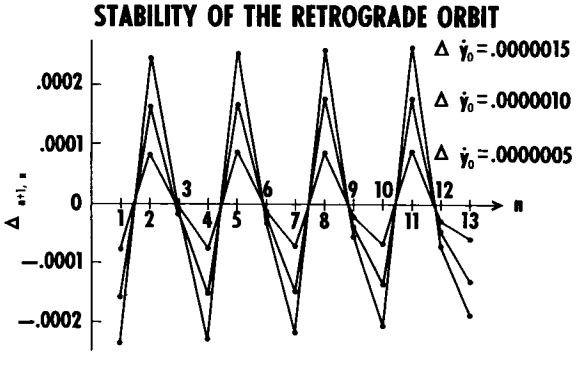


FIGURE 1.—Stability of the retrograde orbit showing the variation of $\Delta_{n+1,n}$ with n. The amplitude of variation appears to be proportional to the deviation $\Delta\dot{y}_0$ of the initial ejection velocity from the correct value. This figure provides a heuristic argument for the existence of periodic orbits in a mathematical sense.

with n. We immediately find that, in the case of *direct* orbits, $\Delta_{n+1,n}$ oscillates with an ever increasing magnitude roughly as the exponential function of n. This clearly indicates instability of the periodicity. For *retrograde* orbits, on the other hand, the variation in $\Delta_{n+1,n}$ with n for each given $\dot{y}$ slightly different from the correct $\dot{y}_0$ of the periodic orbit is simply oscillatory without any increase in magnitude, as figure 1 shows. Moreover, the amplitude of variation in $\Delta_{n+1,n}$ decreases with the decrease in the deviation of $\dot{y}_0$ from the correct value that corresponds to the periodic orbit. This shows most clearly that the periodic orbit is stable under a small change in the initial conditions.

REFERENCES

1. SZEBEHELY, V., "The Restricted Problem of Three Bodies," General Electric Co., Space Sci. Lab., Philadelphia, Pa., 1961.
2. HUANG, S.-S., "Preliminary Study of Orbits of Interest for Moon Probes," *Astronom. J.*, **67**(5): 304–310, June 1962.
3. MESSAGE, P. J., "Some Periodic Oribts in the Restricted Problem of Three Bodies and Their Stabilities," *Astronom. J.*, **64**(1271): 226–236, August 1959.
4. NEWTON, R. R., "Periodic Orbits of a Planetoid Passing Close to Two Gravitating Masses," *Smithsonian Contrib. to Astrophys.*, **3**(7): 69–78, 1959.

THE COMPARISON AND COMBINATION OF SATELLITE WITH OTHER DETERMINATIONS OF GEODETIC PARAMETERS

WILLIAM M. KAULA

Goddard Space Flight Center

Three areas of geodetic interest can be defined in which it is desirable to compare satellite results with those obtained by terrestrial and lunar methods:

1. Variations in the gravity field. Determinations from satellite orbits of tesseral harmonics need an independent check. Terrestrial gravimetry at present is of dubious adequacy, except to confirm the order of magnitude of the variations determined by use of autocovariance analysis. An excellent partial check on the phase as well as the amplitude of the variations is the astro-geodetic geoid. For comparison satellite data must be transformed from a harmonic to a spatial representation.

2. Station positions. Satellite data so far have yielded only the rather negative conlcusion that the position determinations of the major continental geodetic systems are probably correct to within 100 meters.

3. The scale relationship between laboratory and satellite systems, most conveniently expressed in terms of GM. An appreciable discrepancy already exists between the GM value obtained from lunar observations and that obtained from terrestrial geodesy. Thus far, satellite results in this area are inconclusive.

INTRODUCTION

The fullest utilization of satellite geodesy requires the development of methods to compare its results with those obtained by other means. Satellites, after all, are rather limited as measuring devices. Therefore, their results are subject to distortion and independent confirmation is desirable. Such confirmation is difficult; in fact, our interest in satellite geodesy is largely generated by the incompleteness and distortions of other methods.

This paper is a brief commentary on satellite geodesy in relation to conventional methods. We shall discuss this subject from three standpoints: the geoid (or gravitational field); station positions; and the scale relationship between satellite and laboratory measuring systems.

THE GEOID

The results of satellite orbit studies to obtain details of variations in the earth's gravity field are presently somewhat static, because orbital variations which must be used to determine the tesseral harmonics accurately are orders of magnitude smaller than variations used to determine the zonal harmonics. To enable analysis of such variations, significant improvement is needed in at least one of the following respects:

1. Better orbits-perigees above 800 km, eccentricities on the order of 0.01–0.12, and a variety of inclination.

2. Frequent accurate tracking of all parts of the orbit, both day and night sides.

3. Better physical models of the upper atmosphere.

4. Better data analysis techniques.

Items 1 and 2 are already forthcoming, so it is timely to consider what sort of independent checks might be available.

The first and most evident of such checks is the

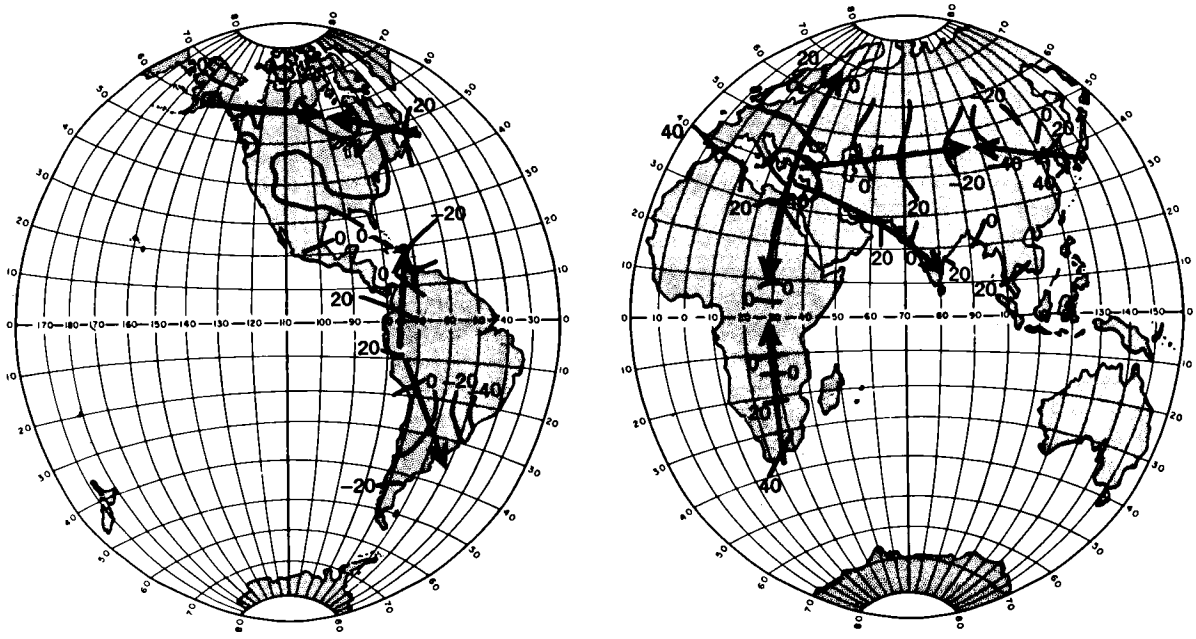


FIGURE 1.—Geoid heights (meters) based on astro-geodetic data (reference 6).

latitude variation observations indicating that the harmonic U_2^1, which would exist only if rotation of the earth was not about a principal axis, is negligibly small. It is thus hoped that satellite orbit analyses incorporate a sufficient variety of conditions that Y_2^1 can be included in the solution; its difference from zero in the results will then be some measure of reliability of the determinations for other harmonics.

Further checks from terrestrial geodesy are the gravimetric and astro-geodetic data. Both are very incomplete. It is doubtful that any existing treatment of gravimetry is adequate: Those of Heiskanen[1] and Kaula[2] depend too much on step-by-step extrapolation and correlation with topography; those of Jeffreys[3] and Zhongolovitch[4] are distorted by approximations used to overcome the poor distribution of observations. The recent results of Uotila,[5] based on much more data, may constitute an appreciable improvement.

The astro-geodetic geoid is good but is limited in coverage. Figure 1 shows the latest geoid

TABLE 1.—*Gravitational Harmonics Obtained from Satellite Orbits*

(multiply by a scaling factor of 10^{-6})

n	m	Conventional $C_{n,m}$	Conventional $S_{n,m}$	Partly Normalized $A_{n,m}$	Partly Normalized $B_{n,m}$	Fully Normalized $A_{n,m}$	Fully Normalized $B_{n,m}$	RMS Geoid Height Contribution (meters)
2	2	2.1	-0.8	10.1	-3.7	3.2	-1.2	±22
3	0			2.30		0.87		6
4	0	1.80		1.80		0.60		4
4	1	-1.2	-0.2	-5.5	-1.0	-1.3	-0.2	8
5	0	0.30		0.30		0.09		±1

published by Fischer.[6] Heavy arrows have been inserted on it to mark all major slopes which might be indicated by satellite orbits—those longer than $\frac{1}{4}$ wavelength ($22\frac{1}{2}°$) of a fourth degree harmonic and having a geoid height change of more than 40 m. Figure 2 shows the geoid corresponding to table 1, which consists of results derived solely from satellite data. The zonal harmonics in table 1 are rounded-off averages of the results obtained by the principal investigators, with greatest weight given to those of Kozai;[7] the $n,m=2,\ 2$ and $4,\ 1$ harmonics are those given by Newton,[1] at present the only published data at least partly satisfying items 1 and 2 on the list given above. $C_{n,m}$, $S_{n,m}$, $A_{n,m}$, and $B_{n,m}$ in table 1 are as defined by the International Astronomical Union recommendation.[9] $\bar{A}_{n,m}$ and $\bar{B}_{n,m}$ are the coefficients of

$$\frac{\mu}{r}\left(\frac{a_e}{r}\right)^n \sqrt{\frac{(n-m)!}{(n+m)!}\,(2n+1)\kappa_m}\,\frac{\cos^m\phi}{2^n\cdot n!}\sum_{t=0}^{\frac{n-m}{2}}\frac{(2n-2t)!}{(n-m-2t)!}\binom{n}{t}(-1)^t\sin^{n-m-2t}\phi\begin{Bmatrix}\cos\\ \sin\end{Bmatrix}m\lambda\ ,$$

$$(\kappa_0 = 1;\ \kappa_m = 2,\ m \neq 0);$$

and they have the advantage, for geophysical purposes, of being identical with the rms amplitudes.

On figure 2 arrows have been drawn in locations corresponding to those on figure 1. The solid arrows indicate agreement, and the dashed arrows disagreement, with the astro-geodetic geoid. The only major feature for which there is good agreement is the hollow in the center of Asia. Some of the discrepancies might be explained by the third degree harmonics which are absent from the satellite geoid, but it is disappointing that there is not better agreement in the north-south direction, in which the zonal harmonics would be expected to prevail. It is disturbing that the dominant term in the satellite geoid, $\bar{A}_{2,2}$, $\bar{B}_{2,2}$, is the most poorly determined: the partial derivative of the longitude of Transit IV-A ($1961\mathit{O}_1$) along its orbit with respect to $\bar{A}_{2,2}$, $\bar{B}_{2,2}$, is about half the derivative with respect to $\bar{A}_{4,1}$, $\bar{B}_{4,1}$.

The satellite geoid does not compare any better with any of the gravimetric geoids which have

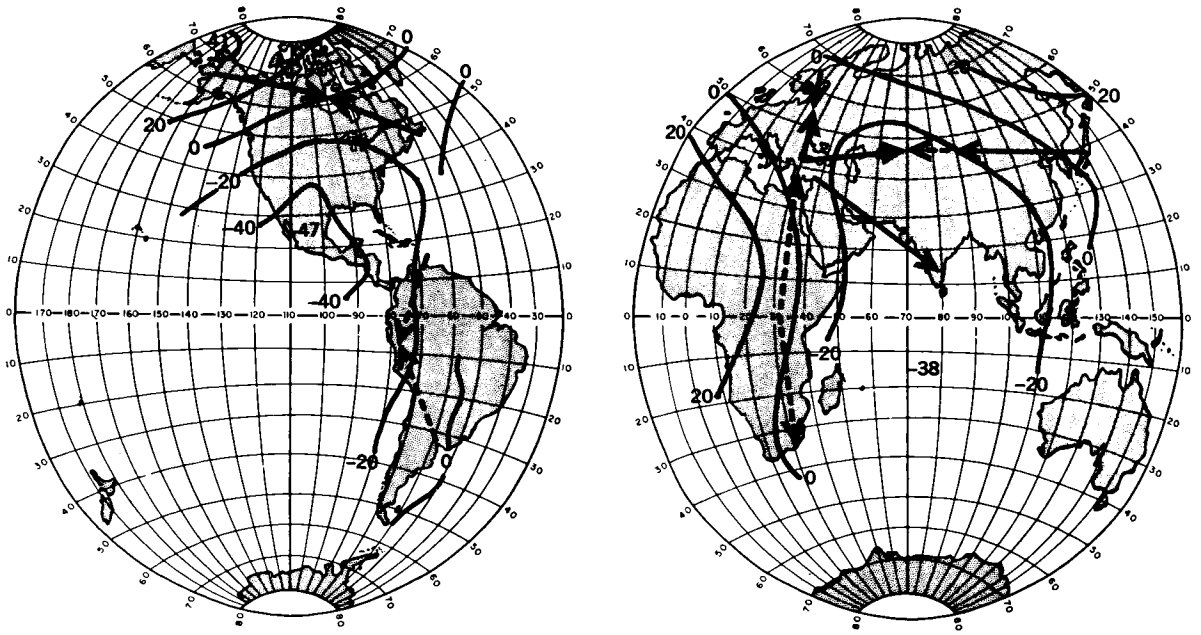


FIGURE 2.—Geoid heights (meters) based on satellite data (references 7 and 8).

TABLE 2—*Comparison of Spherical Harmonics from Satellite Motions with the Order of Magnitude Predicted by Autocovariance Analysis of Gravimetry*

(multiply by a scaling factor of 10^{-6})

n	$\sigma\{\bar{A}_{n,m}, \bar{B}_{n,m}\}$	$\bar{A}_{n,m}, \bar{B}_{n,m}$
3	±1.27	0.87
4	±0.62	0.60, -1.3, -0.2
5	±0.25	0.09

been published. An improvement might be obtained by applying to the gravimetry the generalization of linear autoregression for non-uniformly distributed data. But we do have one limited category of gravimetric data which agrees well—the degree variances from autocovariance analysis, which afford a check on the order of magnitude of the harmonic coefficients of the third and higher degrees. Table 2 compares the magnitude of coefficients predicted by the autocovariance analysis in[2] with those from table 1. Results are not yet sufficiently accurate to make a precise definition of the geoid a matter of concern in any comparisons. An interesting study by Pellinen[10] gives effects on the order of $\pm 0.08 \times 10^{-6}$ on $\bar{A}_{1,m}$, $\bar{B}_{1,m}$ from the higher order term in Molodenskii's theory.

STATION POSITIONS

Since the errors in station positions connected to the major geodetic datums are believed to be smaller than the principal effects of the tesseral harmonics on satellite orbits, determinations for datum shifts so far are even more speculative than those for the gravitational field. Experience has borne out the estimates of station position accuracies fairly well: Usually position shifts of less than 100 m are required to get a best fit to satellite observations for stations connected to the principal geodetic systems of America and of Eurasia-Africa, but shifts of more than 100 m are ordinarily needed for a station on an isolated system such as Hawaii, Australia, or Argentina. The sets of observations have not been well-conditioned enough, however, to give results which could be called conclusive.

SCALE RELATIONSHIP

Despite drag and other complications, the semimajor axis giving a mean motion that fits observations of a satellite over more than a few days is rather accurately determined—usually with a proportionate error of less than 10^{-6}. Holding fixed station positions, or the differences between them, suggests a possibility for determining the scale relationship of the satellite system to the cgs system. This relationship is usually expressed in terms of GM, which ties together time and length through Kepler's law: $n^2a^3 = GM$. There now exists a proportionate discrepancy of 3×10^{-5} between GM from terrestrial data and GM from the lunar mean motion and radar distance measurements. In the terrestrail determination the most likely defect is the equatorial radius as found by fitting the astro-geodetic to the gravimetric geoid; in the astronomic determination, the most likely defect is the lunar mass deduced from the lunar inequality found from observations of Eros.[11] A determination of GM from satellite motions with the station positions held fixed depends on the accuracy of these absolute positions—at best the proportionate error is on the order of 5×10^{-6}, whereas a determination with the *differences* of positions held fixed depends on the accuracy of the triangulation connecting the stations—the proportionate error is on the order of 3×10^{-6} for transcontinental distances. However, neither way can be considered independent of the determination made entirely from terrestrial data.

Determinations from satellite orbits so far have indicated proportionate changes on the order of 10^{-5} in GM—without, however, any marked consistency.

REFERENCES

1. HEISKANEN, W. A., "The Columbus Geoid," *Trans. Amer. Geophys. Union*, **38**(6): 841–848, December 1957.
2. KAULA, W. M., "Statistical and Harmonic Analysis of Gravity," *J. Geophys. Res.*, **64**(12): 2401–2421, December 1959.
3. JEFFREYS, H., "The Determination of the Earth's Gravitational Field (Second Paper)," *Monthly Not. Royal Astronom. Soc. Geophys. Suppl.*, **5**(3): 55–66, 1943.

4. ZHONGOLOVICH, I., "The External Gravity Field of the Earth and the Fundamental Constants Connected with it," *Trudy Instituta Teoreticheskoy Astronomii* (Akademiya Nauk SSSR), No. **3**, 1952.
5. UOTILA, U. H., "Theoretical Gravity Formula Corresponding to Current Gravity Holdings," Paper presented to the 43d Annual Meeting of the American Geophysical Union, Washington, April 1962.
6. FISCHER, I., "An Astrogeodetic World Datum from Geoidal Heights Based on the Flattening $f = 1/298.3$," *J. Geophys. Res.*, **65**(7): 2067–2076, July 1960.
7. KOZAI, Y., "The Gravitational Field of the Earth Derived from Motions of Three Satellites," *Astronom. J.*, **66**(1): 8–10, February 1961.
8. NEWTON, R. R., Presentation to the Internat. Sympos. on the Use of Artificial Satellites for Geodesy (Sponsored by the Committee on Space Research and the International Association of Geodesy), Washington, April 1962.
9. HAGIHARA, Y., "Recommendations on Notation of the Earth Potential," *Astronom. J.*, **67**(1): 108, February 1962.
10. PELLINEN, L. P., "Accounting for Topography in the Calculation of Quasigeoidal Heights and Plumb-Line Deflections from Gravity Anomalies," *Bull. Geod.*, **63**, 57–65, March 1962.
11. RABE, E., "Derivation of Fundamental Astronomical Constants from the Observations of Eros During 1926–1945," *Astronom. J.* **55**(4): 112–126, June 1950.

DETERMINATION OF THE EARTH'S GRAVITATIONAL FIELD

WILLIAM M. KAULA

Goddard Space Flight Center

In all three principal methods of measuring the earth's gravitational field—gravimetry, satellite orbit perturbations, and astrogeodetic networks—both mathematical theory and observational techniques have been developed in recent years to more than sufficient accuracy to define and determine the irregularities of the field significant on a planetary scale. The common defect of all three methods is inadequate distribution of observations. At present, the different methods agree only for major features such as the Indian Ocean minimum, and the normalized spherical harmonic coefficients of the potential $\bar{C}_{nm}$, $\bar{S}_{nm}$ are known to be of an order of magnitude of about $\pm 1.3\times 10^{-6}/(n-1)$ with an uncertainty of about $\pm 0.6\times 10^{-6}$ for the tesseral terms. Gradual improvement is anticipated with additional observations and the application of better statistical techniques made possible by modern computers.

INTRODUCTION

The subject of this review is the determination of the earth's gravitational field, given measurements that are adequate in accuracy but incomplete in their distribution. The emphasis will be on the wide-scale variations of the field rather than on local variations. In accordance with this emphasis, we shall usually employ the spectral representation of the gravitational field rather than the spatial.

Detailed discussion will be attempted only for developments since the advent of artificial satellites in 1957. Earlier developments are described adequately in publications generally available: the texts by Heiskanen and Vening-Meinesz (1958), Jeffreys (1959), and Bomford (1962); the Handbuch der Physik articles by Garland (1956) and Jung (1956); and, recently translated into English, the important treatise by Molodenskiy, Yeremeyev, and Yurkina (1960).

MATHEMATICAL EXPRESSION OF THE GRAVITY FIELD

The Reference Figure

The physically logical reference figure is that of a rotating fluid in equilibrium. The principal variable involved in such a model is the radial density distribution. The theoretical investigation of rotating fluid models of the earth is still being pursued, principally by Ledersteger (1962; see this reference for earlier papers).

The more mathematically tractable ellipsoid of revolution, which differs by quantities of the order of 10^{-6} from a rotating fluid, is generally used in practice for a reference figure. A closed expression for the gravity (i.e., gravitational attraction plus centrifugal force), first obtained by Pizzetti (1894), requires the use of ellipsoidal harmonics. For practical application, the present accuracy of determination of the gravity field makes it more convenient to use series developments in spherical harmonics. For an ellipsoidal figure specified by an equatorial semimajor axis a, an equatorial gravity γ_e, a polar semiaxis b, and a rate of rotation ω, formulas of sufficient precision for the subsequent discussions in this review are (Jung, 1956, pp. 542–562; Heiskanen and Vening-Meinesz, 1958, pp. 51–53; Bomford, 1962, pp. 414–416)

$$f=(a-b)/a \tag{1}$$

$$m=\omega^2 a/\gamma_e \tag{2}$$

$$kM=a^2\gamma_e[1-f+\tfrac{3}{2}m-\tfrac{15}{14}mf+0(f^3)] \tag{3}$$

$$\gamma=\gamma_e[1+(\tfrac{5}{2}m-f-\tfrac{17}{14}mf)\sin^2\phi_g + (\tfrac{1}{8}f^2-\tfrac{5}{8}mf)\sin^2\phi_g+0(f^3)] \quad (4)$$

$$r_e=a[1-(f+\tfrac{3}{2}f^2)\sin^2\phi_g+\tfrac{3}{2}f^2\sin^4\phi_g+0(f^3)] \quad (5)$$

$$J_2=\tfrac{2}{3}f(1-\tfrac{1}{2}f)-\tfrac{1}{3}m[1-\tfrac{3}{2}m-\tfrac{2}{7}f]+0(f^3) \quad (6)$$

$$J_4=-\tfrac{4}{35}f(7f-5m)+0(f^3) \quad (7)$$

$$U=(kM/r)[1-J_2(a/r)^2P_2(\sin\phi) - J_4(a/r)^4P_4(\sin\phi)+0(f^3)] \quad (8)$$

where k is the gravitational constant, M is the mass, ϕ is the geocentric latitude, ϕ_g is the geodetic latitude, r_e is the radial spherical coordinate of the surface of the ellipsoid, U is the gravitational potential, and the P_n are Legendre polynomials. Equations 3 through 7 are essentially those given by Helmert (1884) and are used by most geodesists. Another set derived by DeSitter (1938) is preferred by most astronomers.

The expression of the external field of an ellipsoid of revolution has recently been developed in series to higher-order terms by Cook (1959), Hirvonen (1960), and Lambert (1961) and expressed in closed formulas by Caputo (1963).

Variations in the Field: First Approximation

To accommodate departures of the gravitational potential of the earth from the reference figure, the field can be expressed as a sum of spherical harmonics:

$$V=(kM/r)[1+\sum_{n=2}^{\infty}(a/r)^n \cdot\sum_{m=0}^{n}\bar{P}_{nm}(\sin\phi)\{\bar{C}_{nm}\cos m\lambda+\bar{S}_{nm}\sin m\lambda\}] \quad (9)$$

The gravitational potential V is positive, following the sign convention of astronomy and geodesy. The absence of $n=1$ terms from (9) means that the coordinate origin coincides with the center of mass.

$\bar{P}_{nm}(\sin\phi)$ is the Legendre associated polynomial. The overbar signifies a normalization; the one we shall apply in this review makes

$$\int_0^{2\pi}\int_{-\pi/2}^{\pi/2}\left[\bar{P}_{nm}(\sin\phi)\begin{Bmatrix}\cos\\ \sin\end{Bmatrix}m\lambda\right]^2\cos\phi\,d\phi\,d\lambda=4\pi \quad (10)$$

so that

$$\bar{P}_{nm}(\sin\phi)=\left[\frac{(2n+1)(n-m)!(2-\delta_{0m})}{(n+m)!}\right]^{1/2}\frac{\cos^m\phi}{2^n} \cdot\sum_{t=0}^{k}\frac{(-1)^t(2n-2t)!}{(n-m-2t)!t!(n-t)!}\sin^{n-m-2t}\phi \quad (11)$$

where δ_{0m} is the Kronecker delta, and k is the integer part of $(n-m)/2$.

The coefficients $\bar{C}_{nm}$, $\bar{S}_{nm}$ are independent parameters

$$\bar{C}_{20}=-J_2/5^{1/2}\approx-0.0004842 \quad (12)$$

All other $\bar{C}_{nm}$, $\bar{S}_{nm}$ are $0(10^{-6})$. The magnitude of f (or J_2 or $\bar{C}_{20}$) means that variations of the field referred to the ellipsoidal surface r_e expressed by (5) can be expressed by the $\bar{C}_{nm}$, $\bar{S}_{nm}$ with a proportionate error of only 10^{-3}. The variation most commonly referred to the ellipsoid is the geoid, i.e., the equipotential surface of the family defined by (9) for which V is equal to the U of (8) on the ellipsoidal surface in (5). Expressed in this manner, the geoid height is

$$N=(V-U)/\gamma=T/\gamma \quad (13)$$

The ellipsoidal surface r_e is an equipotential for the gravitational potential U *plus* the centrifugal force potential ψ:

$$\psi=(\tfrac{1}{2}\omega^2)r^2\cos^2\phi \quad (14)$$

Measurements of the gravity acceleration can be referred to the geoid through leveling observations. To compare the observed acceleration g with the acceleration γ on the reference model, we must therefore allow for the height difference N:

$$\Delta g=g(r_e+N)-\gamma(r_e)=-\left[\frac{\partial(V+\psi)}{\partial n}\right]_{r=r_e+N} +\left[\frac{\partial^2(U+\psi)}{\partial n^2}N+\frac{\partial(U+\psi)}{\partial n}\right]_{r=r_e} \quad (15)$$

where the derivations are normal to the ellipsoid.

Using (13) and taking the spherical approximation $-2\gamma/r$ for $\partial\gamma/\partial r$, we obtain

$$\Delta g=(\partial T/\partial n)-2T/r \quad (16)$$

The last term in (16) is usually referred to as Bruns' term.

For the coefficient $\bar{A}_{nm}$, $\bar{B}_{nm}$ of a particular term in the spherical harmonic expression of Δg, we obtain from (9), (15), and (16)

$$\bar{A}_{nm},\ \bar{B}_{nm}=\gamma(n-1)\{\bar{C}_{nm},\ \bar{S}_{nm}\} \tag{17}$$

To relate the geoid height N at a particular point to the anomalies Δg over the surface of the earth, we develop the anomalies in spherical harmonics about the point as a pole and use the orthogonality relationships between spherical harmonics to obtain:

$$N=\frac{R}{4\pi\gamma}\int_{\text{sphere}} S(\cos\theta)\ \Delta g\ d\sigma \tag{18}$$

where

$$S(\cos\theta)=\sum_{n=2}^{\infty}\frac{(2n+1)^{1/2}}{n-1}\bar{P}_n(\cos\theta) \tag{19}$$

$\bar{P}_n(\cos\theta)$ is the normalized zonal harmonic. The closed form of $S(\cos\theta)$ is Stokes' function [Stokes, 1849):

$$S(\cos\theta)=\csc\tfrac{1}{2}\theta-3\cos\theta\ln(\sin\tfrac{1}{2}\theta+\sin^2\tfrac{1}{2}\theta)-6\sin\tfrac{1}{2}\theta+1-5\cos\theta \tag{20}$$

The similar function for the slope of the geoid, or deflection of the vertical, was obtained by Vening-Meinesz (1928) and de Graaff-Hunter (1935).

If the coordinate system were rotated and the generalized additional theorem applied to the spherical harmonics, the coefficients $\bar{C}_{nm}$, $\bar{S}_{nm}$ referred to the new axes would be expressed entirely as functions of the coefficients of degree n referred to the old axes. Furthermore, the quantity

$$\sigma_n^2=\sum_m\{\bar{C}_{nm}^2+\bar{S}_{nm}^2\} \tag{21}$$

remains invariant under rotation (Kaula, 1959b). Thus, the 'degree variance' σ_n^2 is a measure of the amount of variability in a certain wavelength that is independent of the coordinate axes' directions. By summary σ_n^2 over n from 2 to infinity, we obtain the mean square amplitude.

The foregoing relationships were all derived on the assumption that the functions were harmonic, which requires that the geoid be external to all masses. Since the actual geoid is not external to all masses, hypothetical transfers of mass must be made before 'reduction to sea level' of gravity observations and application of Stokes' function. This problem has historically occupied most of the attention of physical geodesy (Heiskanen and Vening-Meinesz, 1958, pp. 147–186; Bomford, 1962, pp. 416–446). Advocates can still be found for most of the reductions that have been proposed in the past. The free air anomaly, which implies that the external topography is condensed in a thin layer just below the geoid; smoothed modifications of the free air anomaly (de Graaff-Hunter, 1960); the isostatic anomaly, which implies that topographic excesses and deficiencies are redistributed uniformly through the crust (Pratt-Hayford compensation) or transferred to the crust-mantle interface (Airy-Heiskanen compensation); and even the Rudzki anomaly, which implies that the external masses are replaced by internal masses yielding the same geoid (Tengstrom, 1962) are all still applied for geodetic purposes, while the Bouguer anomaly, which implies complete removal of the mass excess and deficiencies, is most often applied in geologic interpretation.

The free air anomaly is the most often used, mainly for the practical reason that it does not require detailed evaluation of the topography (except for the local area). As Levallois (1962) emphasized, all the methods of reduction can be considered as equivalent provided that the 'indirect effect' (the effect on Δg of the geoid shift that occurs because of the implied mass transfer) of each method is properly taken into account. Hence the practical distinction becomes which type of anomaly can be considered as most 'representative'; i.e., for which type can the observed point values be interpolated most easily.

Isostatic anomalies are generally regarded as the most representative, but they are also the most laborious, since it is difficult to automate estimation of topographic elevations. Work is continuing to facilitate Airy-Heiskanen isostatic reductions, such as, e.g., the reduction maps for the distant zones of the Isostatic Institute (Kärki et al., 1961). For analytic application of the Airy-Heiskanen isostatic reduction to spherical har-

monics, formulas of Jung (1952) are convenient. The isostatic reduction can be expressed as a power series in $h(\phi, \lambda)$, the height of the solid surface, and $w(\phi, \lambda)$, the depth of the water. By expressing w as a surface coating, Jung develops the isostatic correction in terms of the spherical harmonic expansion of the topography plus water to quadratic terms in the topography of the form:

$$\Delta g_{is} = \sum_{n=0}^{\infty} \{e_n[E]_n + f_n[E^2]_n\} \qquad (22)$$

where $[E]_n$ and $[E^2]_n$ are the sums of terms of degree n in the harmonic expression of

$$E = h - (1.03/\sigma_0)w \qquad (23)$$

and its square, and e_n and f_n are functions of the depth of compensation and the crustal and mantle densities. The functions e_n and f_n contain terms of the form $\{1-(1-t)^{2n+1}\}$, where t is the ratio of the compensation depth to the earth's radius, and so (22) is probably imprecise for higher values of n for this reason as well as because of the lack of higher powers of E. For the purposes of this review, (22) can be considered a means of estimating Δg where E is known but not g.

Variations in the Field: Higher Approximations

As in any problem, further elaboration can be made either in the reference model or in the expressions of departures from the model. The first type of elaboration was made by Sagrebin (1956) who refined Stokes' formula (equation 18) to an accuracy of order f^2 by using elipsoidal harmonics; his solution has since been improved and corrected by Molodenskiy et al. (1960) and by Bjerhammar (1962b). These investigators found that the error in Stokes' formula was indeed of the order of f, i.e., 20 cm or less in N.

The principal theoretical developments in physical geodesy in recent years have been in nonlinear expression of the departures from the model by considering the problem at the physical surface of the earth rather than at the geoid, thus eliminating the inaccuracies due to ignorance of the density within the earth. These developments were initiated in a paper by Jeffreys (1931), and have been carried forward principally by Molodenskiy (1945, 1948). Full details are given by Molodenskiy et al. (1960), and a brief summary by Molodenskiy (1962). The derivation of the method currently applied in the USSR starts with the expression of the disturbed potential T in the form of a surface layer φ on the physical surface of the earth. Taking the derivative of T normal to the physical surface, using the Bruns' equation 16, and allowing for the slope of the surface with respect to the equipotential, we obtain an expression for the free air anomaly Δg_F in terms of integrals over the equipotential surface S. When we assume that the projection of the surface layer φ onto a sphere $\bar{S}$, of radius R, can be expressed in terms of a power series over the discrepancy between $\bar{S}$ and S, the solution is obtained as a successive approximation in four cycles, $n=0$ through 3, over (setting negatively subscripted variables zero)

$$G_n = R\int \frac{h-h_0}{r_0^3}\Big[R\chi_{n-1} - \tfrac{3}{4}(h-h_0)\chi_{n-2} - \tfrac{3}{2}R\left(\frac{h-h_0}{r_0}\right)^2\chi_{n-3}\Big]d\sigma + 2\pi\chi_{n-2}\tan^2\alpha + \delta_{0ri}\,\Delta g_F \qquad (24)$$

$$\chi_n = \frac{G_n}{2\pi} + \frac{3}{(4\pi)^2}\int G_n[S(\cos\theta) - \tfrac{1}{2}]d\sigma \qquad (25)$$

where h_0 is the 'normal' height at the point where the quantity on the left is computed, h is the 'normal' height at the location of the quantities within the integral, r_0 is the chord distance of the sphere between the two points, α is the slope of the surface at the point of computation, and the integration is over the unit sphere. The 'normal' height differs from the orthometric height in that the normal gravity is used in place of the true gravity g. Then to obtain T, the disturbance of the potential:

$$T_n = \frac{R}{4\pi}\int G_n[S(\cos\theta) - \tfrac{1}{2}]d\sigma - \tfrac{1}{2}R^2\int \frac{(h-h_0)^2}{r_0^3}\chi_{n-2}\,d\sigma \qquad (26)$$

$$T = \sum_n T_n \qquad (27)$$

Similar formulas are derived for the deflection of the vertical.

Theoretical developments similar to Molodenskiy's have been made by Levallois (1957), Arnold (1959a, b), de Graaff-Hunter (1960), Hirvonen (1960), Moritz (1961), Bjerhammar (1962a), and Cook (1963b). Summaries and comparisons are given by Arnold (1960a), Hirvonen (1961), Tengstrom (1961), and Molodenskiy et al. (1962), who criticize the approximations made in some of the other theories.

Examples of the numerical significance of the improved theory are given by Molodenskiy et al. (1960), Arnold (1960b), Tengstrom (1961), and Pellinen (1962). The conclusion is that the difference in geoid height is always small, of the order of tens of centimeters at most, but that the effect on the deflection of the vertical can be appreciable, several seconds of arc, in extreme situations. Of particular interest for the determination of the long-wave components of the field is the computation by Pellinen (1962) of the effect of the G_1 term from (24) on the low-degree harmonics. He found that the rms effect up through degree 3 was ± 0.20 mgal on the normalized Δg coefficients, or about $\pm 0.15 \times 10^{-6}$ for the potential coefficients $\bar{C}_{nm}$, $\bar{S}_{nm}$.

STATISTICAL CONSIDERATIONS

Random processes referred to two types of continuums are of concern: time series, which comprise the perturbations of a satellite orbit, and distributions over a spherical surface, which comprise the variations of the earth's gravity field.

We shall first discuss heuristically the general questions of spectral analysis and estimation by quadratic sum minimization, and then apply the conclusions to the particular cases of time series and distributions over a spherical surface.

Spectral Analysis and Quadratic Sum Minimization

Discrepancies $f(s)$ observations from a mathematical model that are small enough to be considered as linear can be represented as a linear transformation of a set of parameters $\mathbf{x}$:

$$f(s) = \mathbf{c}^T(s)\mathbf{x} \qquad s \in S \tag{28}$$

The superscript T denotes the transpose. The coordinates s may be of any number of dimensions. S is a subspace, not necessarily connected, of a total space T. Some of, or all, the parameters $\mathbf{x}$ are members of an enumerable infinite set which are orthogonal over T; i.e., the square array

$$\mathbf{A} = \int_T \mathbf{c}(t)\mathbf{c}^T(t)\, d\tau \tag{29}$$

is diagonal ($d\tau$ is an element of volume in T). The functions $\mathbf{c}(t)$ are usually normalized so that the diagonal elements of $\mathbf{A}$ are either 1 or a constant times $\int_\tau d\tau$. $\mathbf{x}$ can be divided into subsets $\mathbf{x}_n$ such that $\mathbf{c}_n{}^T(t)\mathbf{x}_n$ is invariant under rotation of the coordinate systems. The quadratic sum

$$\sigma_n{}^2 = \mathbf{x}_n{}^T\mathbf{x}_n \tag{30}$$

over one of these subsets is the degree variance, already mentioned in (21) for the case of T which is a spherical surface. For many phenomena, only the $\sigma_n{}^2$ are of interest, and not the individual components of $\mathbf{x}_n$. Usually it is assumed that the random process is isotropic, i.e., the covariance

$$K(r, s) = E\{ f(r)f(s) \} \tag{31}$$

where E denotes the mean value over T and is a function of only the distance (the length along the geodesic) between r and s. Under isotropy, the covariance can be expressed as

$$K(r, s) = \sum_n k_n c_{n0}(d_{rs})\sigma_n{}^2 \tag{32}$$

where d_{rs} is the distance between r and s, c_{n0} is a member of $\mathbf{c}_n$, and k_n is constant. The process of spectral analysis is that of forming numerical estimates of $K(r, s)$ and then using the orthogonality expressed by (29) to determine the $\sigma_n{}^2$. The principal problem in this process is sampling: (1) the distribution of sample points, considering the wavelengths corresponding to the highest degree $\sigma_n{}^2$ anticipated to be significant (the "aliasing" problem); and (2) the weighting, if any, applied to take into account the incomplete extent of the observed subspace S with respect to the total space T (the "window" problem).

If the individual components of $\mathbf{x}_n$ are wanted, or if some of the parameters of $\mathbf{x}$ are not coefficients of orthogonal functions, then spectral

analysis is insufficient and quadratic sum minimization must be applied. In this case, the estimates $K(r, s)$ are still made, but the $\mathbf{x}$ are determined so that, subject to (28),

$$\int_S \int_S \mathbf{x}^T \mathbf{c}(r) K^{-1}(r, s) \mathbf{c}^T(s) \mathbf{x}\, dr\, ds = \text{minimum} \quad (33)$$

where the inverse $K^{-1}(r, s)$ is a function such that

$$\int_S K^{-1}(r, s) K(s, t) ds = \delta(r, t) \quad (34)$$

where $\delta(r, t)$ is the Dirac delta function. The solution is

$$\mathbf{x} = \mathbf{W} \int_S \int_S \mathbf{c}(r) K^{-1}(r, s) f(s) dr\, ds \quad (35)$$

where $\mathbf{W}$ is a symmetric matrix whose elements are the estimated variances and covariances of the χ's: e.g., σ_n^2/N_n for the orthogonal function coefficients, where N_n is the number of terms of degree n.

The minimization of (33) is most generally proved as the determination of the projection $\mathbf{c}^T\mathbf{x}$ in the subspace specified by the chosen elements of $\mathbf{x}$ of the vector $f(r)$ in a Hilbert space with metric $K^{-1}(r, s)$, or measure $K(r, s)$.

More familiar is the form of (28) and (33) where (1) $f(s)$ is evaluated at a finite set of points, and (2) for some of the parameters in $\mathbf{x}$ there is no a priori estimate of the variances. Also the covariance matrix $\mathbf{W}$ of the $\mathbf{x}$ may be known rather than $\mathbf{K}$, the matrix equivalent of $K(r, s)$. In this case, the set of points can be expressed as a vector, and the varianceless part of $\mathbf{x}$ broken off as a separate vector $\mathbf{z}$:

$$\mathbf{Cx} + \mathbf{Mz} = \mathbf{f} \quad (36)$$

$$(\mathbf{f} - \mathbf{Mz})^T \mathbf{K}^{-1} (\mathbf{f} - \mathbf{Mz}) = \mathbf{x}^T \mathbf{W}^{-1} \mathbf{x} = \text{minimum} \quad (37)$$

At least one element is nonzero in each row of $\mathbf{C}$.

Solution by the method of Lagrangian multipliers obtains:

$$\mathbf{x} = \mathbf{WC}^T\mathbf{K}^{-1}[\mathbf{I} - \mathbf{M}(\mathbf{M}^T\mathbf{K}^{-1}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{K}^{-1}]\mathbf{f} \quad (38)$$

$$\mathbf{z} = (\mathbf{M}^T\mathbf{K}^{-1}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{K}^{-1}\mathbf{f} \quad (39)$$

where

$$\mathbf{K} = \mathbf{CWC}^T \quad (40)$$

and $\mathbf{I}$ is the identity matrix. The covariance matrix of $\mathbf{z}$ is

$$\mathbf{V} = (\mathbf{M}^T\mathbf{K}^{-1}\mathbf{M})^{-1} \quad (41)$$

and of the corrected $\mathbf{x}$

$$\mathbf{U} = \mathbf{W} - \mathbf{WC}^T\mathbf{K}^{-1} \times [\mathbf{I} - \mathbf{M}(\mathbf{M}^T\mathbf{K}^{-1}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{K}^{-1}]\mathbf{CW} \quad (42)$$

The rows of $\mathbf{WC}^T$ (and columns of $\mathbf{CW}$) in (38) and (42) need be only those of the elements of $\mathbf{x}$ of interest. If $\mathbf{K}$ is calculated by (40), however, the $\mathbf{W}$ must include all parameters of appreciable effect. Conventionally, the elements of $\mathbf{x}$ are sometimes referred to as "observations" and the elements of $\mathbf{z}$ as "parameters"; mathematically, the distinction is whether or not covariance $\mathbf{W}$ can be preassigned.

Sometimes there are added to the system of (36) and (37) "side conditions":

$$\mathbf{Nz} = \mathbf{k} \quad (43)$$

is a minor complication that can be removed by eliminating a corresponding number of elements in $\mathbf{z}$.

Ordinary least squares is the case in which

$$\mathbf{C} = \mathbf{W} = \mathbf{I} \quad (44)$$

Prediction and interpolation by linear regression is the case in which it is desired to estimate f in the subspace $T - S$:

$$f(r) = \mathbf{c}^T(r)\mathbf{x} \qquad r \in T - S \quad (45)$$

or, letting $\mathbf{g}$ be a set of values of f in $T - S$ and $\mathbf{D}$ be the corresponding array of $\mathbf{c}^T$'s, and using (38) (assuming no z)

$$E\{\mathbf{g}\} = \mathbf{DWC}^T\mathbf{K}^{-1}\mathbf{f} = \mathbf{Bf} \quad (46)$$

A result that is also arrived at through solving the Wiener–Hopf equations for the regression coefficient $\mathbf{B}$ is

$$\mathbf{DWC}^T = \mathbf{BCWC}^T$$

$$\mathbf{K}_{gf} = \mathbf{BK}_{ff} \quad (47)$$

Sometimes the solution for $\mathbf{x}$ or $\mathbf{z}$ is made in stages, with different condition equations 36 in each stage, either because the nature of the process is evolutionary (e.g., orbit predictions) or because a single stage computation would be too

unwieldy. In this case, the vector $\mathbf{x}$ for a particular stage can be considered as consisting of three parts: (1) new observations $\mathbf{x}_x$, with associated coefficient matrix $\mathbf{C}_x$ and covariance matrix $\mathbf{W}_x$; (2) old observations $\mathbf{x}_y$, with associated coefficient matrix $\mathbf{C}_y$ and covariance matrix $\mathbf{W}_y$:

$$\mathbf{W}_y = \mathbf{P}_y \mathbf{U}_0 \mathbf{P}_y^T \tag{48}$$

where $\mathbf{U}_0$ is the $\mathbf{U}$ computed by (42) in the previous stage, or a submatrix thereof, and $\mathbf{P}_y$ is a propagation or transition matrix taking into account change of the parameters (or "state variables") specifying the reference model; and (3) previously estimated parameters $\mathbf{x}_z$, with associated coefficient matrix $\mathbf{C}_z$ and covariance matrix $\mathbf{W}_z$:

$$\mathbf{W}_z = \mathbf{P}_z \mathbf{V}_0 \mathbf{P}_z^T \tag{49}$$

where $\mathbf{V}_0$ is the $\mathbf{V}$ of (41) or a submatrix thereof. The $\mathbf{C}_x$, $\mathbf{C}_y$, $\mathbf{C}_z$ are combined to form a new $\mathbf{C}$ and the $\mathbf{W}_x$, $\mathbf{W}_y$, $\mathbf{W}_z$ are combined to form a new $\mathbf{W}$ for another solution according to (36) through (41). In calculating the residuals $\mathbf{f}$ for each new stage, the corrections $\mathbf{x}$, $\mathbf{z}$ from previous stages are incorporated in the reference model.

The foregoing discussion is a consequence of attempts to combine ideas of time series analysis of Bartlett (1956) and Parzen (1961) with least squares and its generalization as given by Arley and Buch (1950) and Brown (1955, 1957). The results can probably be found to be subsumed in the general treatments of Bochner (1955) and Yaglom (1961) by those able to cope with the recondite mathematics therein. The derivation of (36) through (42) has recently been discussed in detail by Stearn and Richardson (1962). The linear regression prediction of (45) through (47) is applied by Moritz (1962b) to observations on a plane and by Parzen (1961) to time series in continuous, rather than discrete, form. Special cases of the staged computation of $\mathbf{x}_x$, $\mathbf{x}_y$, $\mathbf{x}_z$ with $\mathbf{W}_y$, $\mathbf{W}_z$ calculated as in (48) and (49) are: (1) the optimal prediction of Kalman (1960): $\mathbf{W}_z$, $\mathbf{C}_z$, and $\mathbf{M}$ are $\mathbf{0}$; and (2) the Bayes estimation of Parzen (1962a) and the preassigned covariance of Kaula (1961c): $\mathbf{W}_y$ and $\mathbf{C}_y$ are $\mathbf{0}$; $\mathbf{P}_z$ is $\mathbf{I}$; and $\mathbf{C}_x$, $\mathbf{C}_z$, $\mathbf{M}$, $\mathbf{f}$ have the forms

$$\mathbf{C}_x = \{\mathbf{C}_x/\mathbf{0}\} \qquad \mathbf{C}_z = \{\mathbf{0}/-\mathbf{I}\}$$
$$\mathbf{M} = \{\mathbf{M}_x/\mathbf{I}\} \qquad \mathbf{f} = \{\mathbf{f}/\mathbf{0}\} \tag{50}$$

Another variation of estimation by quadratic sum minimization is Gram–Schmidt orthogonalization (e.g., Robinson, 1959). The function $f(s)$ of (28) is represented as

$$f(s) = \mathbf{d}^T \mathbf{y} \qquad s \in S \tag{51}$$

where the functions $\mathbf{d}$ are orthogonal over S, not T; i.e.,

$$\int_S \mathbf{d}(s)\mathbf{d}^T(s)\,d\tau = \mathbf{I} \tag{52}$$

the identity matrix. Setting

$$\mathbf{B}\mathbf{d} = \mathbf{c} \tag{53}$$

we get

$$\mathbf{B}\mathbf{B}^T = \int_S \mathbf{c}\mathbf{c}^T\,d\tau \tag{54}$$

$\mathbf{B}$ is thus not unique. The Gram–Schmidt orthogonalization process finds a unique $\mathbf{B}$ by requiring that, for the arbitrarily selected sequence of functions in $\mathbf{c}$, $\mathbf{B}$ is triangular. Letting the subscripts on d_i, c_i denote ordering in the arbitrary sequence Gram–Schmidt orthogonalization yields

$$d_1(s) = c_1(s) \tag{55}$$

$$d_i(s) = c_i(s) - \sum_{k=1}^{i-1} \left(d^k(s) \int_S d_k c_k\,d\tau \Big/ \int_S d_k^2\,d\tau \right) \qquad i > 1 \tag{56}$$

If the $\mathbf{x}$ of (28) has a finite number of members, $\mathbf{x}$ can be determined by simple least squares; i.e., in (39) $\mathbf{z}$ represents $\mathbf{x}$, $\mathbf{K}$ is $\mathbf{I}$, and $\mathbf{M}$ is the array of values of $\mathbf{c}$ and $\mathbf{f}$ is the array of values of f for a set of points or volume element means. In practice, $\mathbf{x}$ has an infinite number of members, and it is not known beforehand which are significant; also, the disposition of S within T may be such that $\mathbf{M}^T\mathbf{M}$ approaches singularity, so that it is computationally impracticable to invert. If Gram–Schmidt orthogonalization is applied, then by (52) $\mathbf{M}^T\mathbf{M}$ is diagonal and (39) degenerates to a simple Fourier analysis in which the coefficients are determined one by one until it is decided that nothing more of significance is being found. Gram–Schmidt orthogonalization thus is a systematic method of estimation that avoids the inversion $(\mathbf{M}^T\mathbf{M})^{-1}$ of simple least squares or the

inversion $\mathbf{K}^{-1}$ of quadratic sum minimization (equation 35 or 38 with $\mathbf{M}=\mathbf{0}$). It does not, however, abolish the "window" effect—the distortion of estimates of lower-degree coefficients by higher-degree variations—and by avoiding the assumptions of stationarity and isotropy of spectral analysis and quadratic-sum minimization it is accordingly a less effective predictor of the variations in the subspace $T-S$.

A technique similar to Gram–Schmidt orthogonalization in principle is step-wise least squares: the parameters in $\mathbf{x}$ are determined one each or one subset at a time, and the contribution of a particular set is subtracted from $\mathbf{f}$ before determining those of the next set. This method is commonly employed in excessively large problems when the parameters are not coefficients of orthogonal functions. It could perhaps be considered as a special case of the staged least squares (equation 50) in which $\mathbf{W}_z$ is set zero, but differs in that the same condition equations 36 are used repeatedly with different sections of the $\mathbf{M}$ matrix set zero, which implies that the situation must approach orthogonality for the procedure to be justifiable.

Time Series

Probability models of time series, and the analysis of continuous or uniformly spaced observations thereof, have been developed in considerable detail (Bartlett, 1956; Grenander and Rosenblatt, 1957; Blackman and Tukey, 1959; Parzen, 1961). Applications to geophysics have been discussed by Holloway (1958), Munk and MacDonald (1960), and Van Isacker (1961).

The time series of significance in the present problem are vector quantities characterized by a discrete spectrum of fixed phase and amplitude plus a continuous spectrum of noise (Kaula, 1963a):

$$x_f(t) = \Re\left\{\sum_{n=1}^{N}\alpha_{fn}\exp[i\lambda_n t]+\int_0^\infty \beta_f(\omega)\exp[i\omega t]\,d\omega\right\} \tag{57}$$

where the subscript f denotes a vector component and n denotes one of a set of frequencies; $\Re$ denotes the real part, i is $(-1)^{1/2}$, and the vectors α_{fn}, $\beta_f(\omega)$ are complex. The λ_n are all known; the α_{fn} are linear functions of a set of parameters p_l, less than N in number, which it is our problem to determine. The duration T of observation of the time series is such that the discrete spectrum stands out above the noise; i.e.,

$$\beta_f(\omega)\,(2\pi/\omega T)\ll\alpha_{fn} \qquad \omega\geq\lambda_n \tag{58}$$

but

$$\beta_f(\omega)\,(2\pi/\omega T)\gg\alpha_{fn} \quad \text{some} \quad \omega\ll\lambda_n \tag{59}$$

If the $x_f(t)$ were continuously observed with observational errors random, the condition expressed by (58) would assure that the p_l could be accurately determined. In the problem of interest, however, the $x_f(t)$ are incompletely and intermittently observed, and furthermore the observations are affected by unknown biases; i.e., $x_f(t)$ must be replaced by

$$x_f(t)\rightarrow I(t)\frac{\partial x_f}{\partial y_j}\left[y_j(t)-\frac{\partial y_j(t)}{\partial q_m}q_m\right] \tag{60}$$

where $I(t)$ is unity during observation and zero at other times, and the number of components denoted by j is less than those denoted by f. Under these circumstances, the effective orthogonality expressed by (58) is destroyed and the condition expressed by (59) greatly increases the length of record required to estimate the parameters p_l, q_m by ordinary methods such as simple least squares, as well as increasing the chance that sources of systematic distortion could be hidden in the noise $\beta_f(\omega)$.

The problem of spectral analysis of time series with missing observations has been treated by Jones (1962) and Parzen (1962b). These discussions treat $I(t)$ as an "amplification factor":

$$x(t) = I(t)\,y(t) \tag{61}$$

and show that for covariance estimates for any lag v

$$\hat{R}(v) = R_x(v)/R_I(v) \tag{62}$$

The condition expressed by (59) suggests, however, that the time series itself would be useful to determine only the covariance due to the low-frequency continuous part of the spectrum, and that since the $\partial\alpha_{fn}/\partial p_l$ are known it would be

better to obtain estimates of the high-frequency discrete spectrum effects from independent estimates of the variances of the p_l's. The practical problem then becomes the large dimension of the covariance matrix which must be manipulated in computation.

Distributions Over a Spherical Surface

Spectral representation on a spherical surface is discussed by Schoenberg (1942), Obukhov (1947), Kaula (1965b), and Jones (1963). In this case, the coordinates s of (28) are latitude and longitude (ϕ, λ) and the functions $c^T(s)$ are surface spherical harmonics. With the normalization specified by (10) and (11), and the assumption of isotropy, the covariance of (32) becomes

$$K(r, s) = \sum_n [\sigma_n^2/(2n+1)^{1/2}]\bar{P}_{n0}(\cos\theta_{rs}) \quad (63)$$

The distributions over a sphere significant in the present problem are scalar quantities characterized by spectrums with significant contributions from wave numbers as high as 4000. Hence the sample to determine the covariances $K(\theta_{rs})$, and thence to estimate the σ_n^2's for small n's should be large and randomly distributed to avoid aliasing (Shapiro and Silverman, 1960). Furthermore, in (38) and (39) the matrix **K** employed must be calculated directly from the estimated covariances $K(\theta_{rs})$ and not from **W**, as stated in (40).

The cross covariance between two different variables f, h on a spherical surface, assuming isotropy, will be

$$K_{xy}(\theta_{rs}) = \sum_n [\sigma_n(fh)/(2n+1)^{1/2}]\bar{P}_n(\cos\theta_{rs}) \quad (64)$$

where

$$|\sigma_n(fh)| \leq [\sigma_n^2(f)\sigma_n^2(h)]^{1/2} \quad (65)$$

If f is known over part (S) of the sphere, and h is known over all (T) the sphere, the best estimate of a coefficient $\bar{C}_{nm}(f)$ will be the weighted mean of the estimate obtained from the f in S by (35) and the estimate obtained using $\bar{C}_{nm}(h)$ and the cross-degree variance $\sigma_n(fh)$:

$$E\{\bar{C}_{nm}(f)\} = (\rho_s+\rho_h)^{-1}\left[\rho_s\frac{\sigma_n^2(f)}{2n+1}\int_S\int_S c_{nm}(r)K^{-1}(r, s)f(s)dr\,ds+\rho_h\frac{\sigma_n(fh)}{\sigma_n^2(h)}\bar{C}_{nm}(h)\right] \quad (66)$$

$$\rho_s = [\sigma^2\{\bar{C}_{nm}(f) \mid f\in S\}]^{-1}$$

$$=\left[\frac{\sigma_n^2(f)}{2n+1}\left\{1-\frac{\sigma_n^2(f)}{2n+1}\int_S\int_S c_{nm}(r)K^{-1}(r, s)c_{nm}(s)dr\,ds\right\}\right]^{-1} \quad (67)$$

$$\rho_h = [\sigma^2\{\bar{C}_{nm}(f) \mid \bar{C}_{nm}(h)\}]^{-1} = \left[\frac{1}{2n+1}\left\{\sigma_n^2(f)-\frac{[\sigma_n(fh)]^2}{\sigma_n^2(h)}\right\}\right]^{-1} \quad (68)$$

Similarly, for the optimum estimate of f at a set of points in $T-S$ a modification of (46), (47) can be used:

$$E\{\mathbf{g}\} = [\mathbf{V}_{gf}^{-1}+\mathbf{V}_{gh}^{-1}]^{-1} \times [\mathbf{V}_{gh}^{-1}\mathbf{K}_{gf}\mathbf{K}_{ff}^{-1}\mathbf{f}+\mathbf{V}_{gh}^{-1}\mathbf{K}_{gh}\mathbf{K}_{hh}^{-1}\mathbf{h}] \quad (69)$$

where

$$\mathbf{V}_{gf} = \mathbf{U}_{gg}-\mathbf{K}_{gf}\mathbf{K}_{ff}^{-1}\mathbf{K}_{fg} \quad (70)$$

$$\mathbf{V}_{gh} = \mathbf{U}_{gg}-\mathbf{K}_{gh}\mathbf{K}_{hh}^{-1}\mathbf{K}_{hg} \quad (71)$$

in which

$$\mathbf{U}_{gg} = \mathbf{I}\sigma^2\{f\} \quad (72)$$

If it is assumed that f can be estimated from h by some physical rule such as the isostatic rule of (22), then the foregoing statistical treatment should be applied to f less that part of f accounted for by the physical rule. This treatment is to be emphasized if the parameters of the rule were determined from samples of extreme, rather than average, characteristics, as were those of the isostatic reductions (Heiskanen and Vening-Meinesz, 1958, pp. 187–221).

Spherical harmonic analysis is most common in geomagnetism, where the standard methods have been to interpolate to a latitude and longitude grid, to fit Fourier series to parallels, and then to

fit Legendre functions to the parallel coefficients (Chapman and Bartels, 1940). Fougere (1963), determines polynomials by Gram–Schmidt orthogonalization and then determines the spherical harmonics corresponding to those polynomials found significant. The methods applied in geomagnetism, bypassing the covariance matrix, seem possible, however, only because of the relatively small contribution of higher harmonics.

USE OF GRAVIMETRY

Observing System

A review of gravimetric techniques will appear shortly (LaCoste and Harrison, 1964; see also Harrison, 1962). Of principal concern in the present review are the distribution of gravimetric data and the magnitude of possible systematic errors.

The distribution of available gravimetry is shown in figure 1, which is based on the map of Uotila (1961) updated by information from Worzel et al. (1963) and McCahan (1963). The principal defect of the distribution has always been, of course, the correlation with topography, both on the wide scale shown by figure 1 and locally, as, for example, in observations being conducted mainly in the valleys of mountainous areas and on the islands in archipelagoes. The principal difficulty in improving worldwide coverage is the development of economic airborne gravimeters with accurate velocity determination or seaborne gravimeters capable of obtaining accurate results in a reasonable rough sea.

Systematic errors possibly affecting series of observations which either are the principal data for an appreciable area or provide the reference or calibration net for other observations include: (1) connection of national and other gravimetric systems to the world reference network; (2) calibration: the scale factors utilized for gravimeters; (3) referral of seaborne measurements to land reference stations; (4) navigation errors affecting the position and Eötvos correction for airborne and seaborne gravimetry; and (5) horizontal acceleration and level error effects on airborne and seaborne gravimetry. Errors in categories (1), (2) and (3) are discussed by Woollard (1961); Woollard and Rose, (1963). Woollard (1961) concludes that the principal gravimetric systems are connected with an average error of about ± 0.4 mgal on the basis of gravimeter connections since 1948 and pendulum connections since 1953. Morelli (1963) is more pessimistic, citing as an example a range of 3.2 mgal in values obtained for the Rome reference station by various observers since 1960. The principal improvement

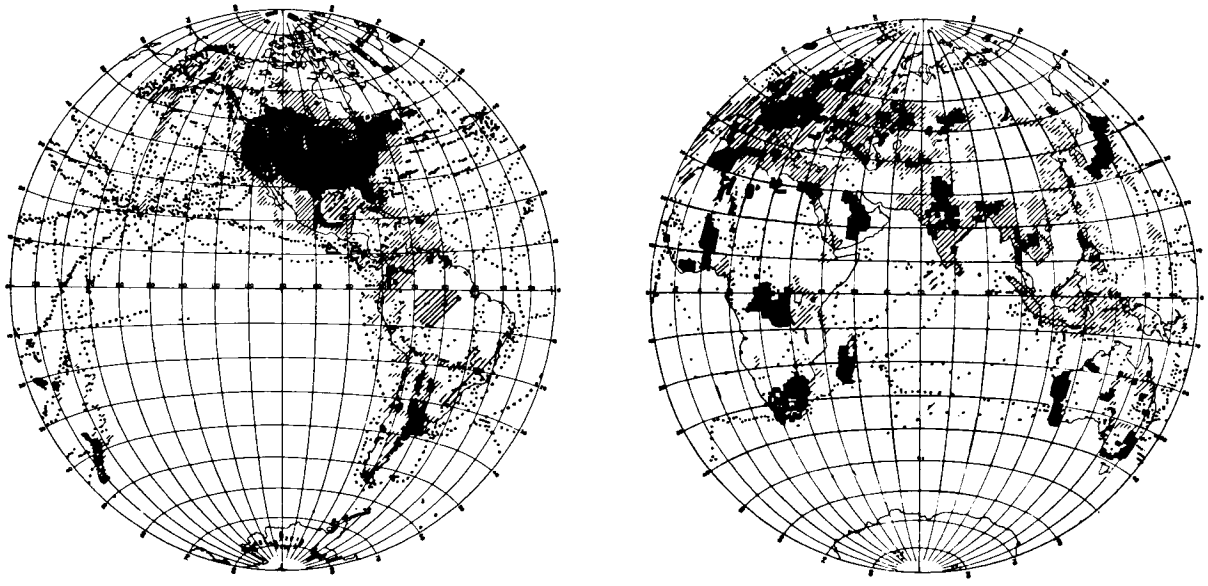


FIGURE 1.—Distribution of available gravimetry.

needed is more reliable north-south pendulum lines to calibrate gravimeters.

The referral of measurements at sea to land reference systems appears to be poorer: Woollard (1961) concludes that Vening–Meinesz's pendulum values are probably reliable to ±2 mgal through measurements made on land with the pendulums, even though 13 of 24 comparisons with Vening–Meinesz's harbor measurements differ by 5 or more mgal. Comparisons from track crossings of different cruises are somewhat better, usually averaging ±5 mgal, which includes other effects. Navigation error is generally considered to be the principal source of error in air and sea measurements, limiting their random accuracy to ±3 to ±5 mgal (Harrison, 1962). There does not seem to be any estimate of systematic navigation error. Horizontal acceleration and level errors dependent on the heading of the ship relative to the sea of several milligals have been reported (Allan et al., 1962). Serious systematic effects of this sort, however, must be less than the ±5 mgal of crossing comparisons with submarine measurements avoiding sea swell effects.

It appears that the most likely source of significant systematic error in gravimetry is in the connection to the reference system of a cruise of marine measurements which constitute the sole data in an appreciable area, but that this error is less than 5 mgal.

The absolute value of gravity is needed for comparison of results obtained from gravimetry with those from satellite orbits. The correction to the Potsdam standard is currently estimated to be about -13 ± 1 mgal. Improvement is anticipated from several determinations in progress (Cook, 1963a).

An important practical task is the collection and processing of the millions of gravimetric observations that have been made. The leader in accomplishing this task has been W. A. Heiskanen, who established the gravity centers at the Isostatic Institute in Helsinki and the Ohio State University in Columbus. Descriptions of the collection and processing are given by Uotila (1960) Heiskanen (1962). Other collections of gravity data exist at the U. S. Naval Oceanographic Office in Washington and the Bureau Gravimetrique Internationale in Paris. Standardization of punched card format for single observations is being worked out by these centers. To exploit fully the existing gravimetry for determination of the worldwide field there is still needed agreement on a standardization of the local treatment to obtain area means for, say, 1° by 1° or 100-km by 100-km blocks, as described in the next section, and of the recording thereof on punched cards or other automated form.

Local Treatment

Gravity anomalies calculated at observation points cannot be used directly in the determination of covariances, harmonic coefficients, and geoid heights as described in preceding sections of this review because of the needs to smooth out the high amount of local variability; to remove, as much as possible, the effect of correlation of observation distribution with local topography; and to keep the computations to a manageable size. The first two of these desiderata indicate the use of a gravity reduction which takes into account the topography. However, only Tanni (1948) and Uotila (1962) have utilized isostatic anomalies. All other calculations of the geoid from gravimetry have placed greater emphasis on the third of the desiderata and have applied at most a simple linear correlation formula to reduce free air gravity anomalies to the mean elevation of the area they represent:

$$\Delta g_m = a_0 + b\bar{h} \tag{73}$$

where $\bar{h}$ is the mean elevation. Hirvonen (1934) and Dubovskiy (Molodenskiy, 1945) did not apply any such correction in their early determinations; Jeffreys (1943) and Zhonglovich (1952) applied the correction to the anomalies for 10° by 10° squares, which were the direct mean of observed values in 1° by 1° squares; Heiskanen (1957), Kaula (1959a, b), and Uotila (1962) used mean anomalies of 1° by 1° squares corrected to mean elevation and then applied step-by-step, or Markov, extrapolation and interpolation to obtain the mean values of 5° by 5° squares. The formation of the mean anomalies for 1° by 1° squares was done largely by Uotila (1960), who found mean values of b of 0.118 mgal/meter on land and 0.069 mgal/meter at

sea. In cases where the distribution of observations with respect to elevation in a 1° by 1° square was insufficient to determine a gradient, Uotila used the gradients from nearby squares or the mean gradients of all squares.

Calculations of isostatic anomalies for much of the available gravimetry have been made at the Ohio State University, the Isostatic Institute (Heiskanen, 1962), and the Bureau Gravimetrique Internationale, but published maps of the reduced anomalies have been limited to a few areas such as Europe and North Africa (Coron, 1962).

The development of computers encourages the application of analytic techniques to replace graphical methods for smoothing, averaging, etc. There have been several papers on use of Fourier series (Tsuboi, 1959); (Bullard and Cooper, 1948); (Tomoda and Ati, 1955); (Dean, 1948) and polynomials (Oldham and Sutherland, 1955); (Brown, 1956); (Grant, 1957); (Krumbein, 1959); (Grant and Elsaharty, 1962); (Mandelbaum, 1963). However, most of these have been devoted to the problem of "continuation" (Peters, 1949); (Hirvonen, 1952); (Tengstrom, 1959); (Orlin, 1959); (Strakhov, 1962) (extrapolation upward, to match airborne measurements, or downward, to determine crustal densities) and assume a fairly good distribution of observations. The most extensive application of Fourier methods has been by Kivioja (1962), who applied them to gravity anomalies over areas up to 10° by 35°. He found distance over which Fourier predictions could be extrapolated to be less than 4°.

The most important application of gravity anomalies has been, and probably will continue to be, the determination of crustal structure in conjunction with seismic and geologic data; e.g., Woollard, (1959); Woollard et al., (1960); Press (1960); Talwani et al., (1961); Oliver et al., (1961). This circumstance suggests that the geology and seismology of an area can be used to predict gravity anomalies where gravimetry is lacking. These geologic methods have been applied most extensively by Durbin (1961) and Woollard (1962). For the south central United States, where the rms anomaly is ±21 mgal, Durbin (1961) reports startlingly good results: the rms error of such predictions is only ±10 mgal. The situation does not appear to be so clear-cut in the more

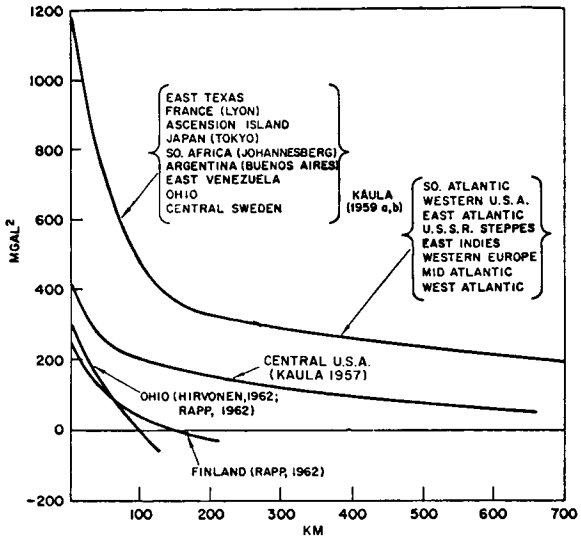


FIGURE 2.—Estimates of local and regional covariance of gravity anomalies.

voluminous study by Woollard (1962), particularly for less stable regions. Also, one suspects that in many areas lacking gravimetry the geologic mapping will also be inadequate.

Statistical analysis has been applied to local variations of gravity anomalies by de Graaf-Hunter (1935), Hirvonen (1956, 1962), Kaula (1957, 1959a, b), Baussus (1960, 1961), Moritz (1962a, b, c), and Rapp (1962). Fundamental to these analyses is the covariance function, as defined by (31). The principal defect of these analyses is the lack of adequate samples (as well as an adequate sampling theory) to obtain numerical estimates of covariance. The largest sample yet analyzed was by Kaula (1959a, b), consisting of nine areas each about 220 km square in size and containing 52 to 140 members. Taking the data in blocks, rather than long lines, severely limits the number of independent estimates, because of the high correlation that will exist between two product pairs which are approximately parallel. The results obtained are shown in figure 2, together with those obtained by Kaula (1957) from profiles across the United States and Hirvonen (1962) and Rapp (1962) from samples in Ohio and Finland, which unfortunately appear to be areas of smaller than average anomalies. Hirvonen finds from his sample that the formula

for covariance over distance d,

$$K(d) = K_0/(1+c^2 d^2) \tag{74}$$

where K_0 and c are arbitrary constants, fits quite well, which is convenient for analytical development of other functions. However, a negative exponential form $K_0 \exp\{-bd\}$ or lower exponent on d in (74) appears to fit better to the larger sample of Kaula (1959a).

More commonly used than the covariance in earlier studies is the mean square anomaly for a square of side length s, related to the covariance K_d by Hirvonen (1962):

$$G_s^2 = \int_0^{\sqrt{2}} W K_d \, dr \tag{75}$$

where

$$r = d/s$$

$$W = (2\pi - 8r + 2r^2)r \quad 0 < r < 1 \tag{76}$$

$$W = [2\pi - 4 - 2r^2 + 8(r^2-1)^{1/2} - 8\tan^{-1}(r^2-1)^{1/2}]r \quad r > 1 \tag{77}$$

If a single observation at the center is taken as representative of a square of side length s, it will have a mean square error of representation:

$$E_s^2 = G_0^2 - G_s^2 \tag{78}$$

The rigorous method of extrapolating and interpolating anomalies is to use the linear regression coefficients as specified by (46) and (47) (Baussus, 1960; Moritz, 1962b). Moritz also generalizes the linear regression to include the topography and anomalies corrected for linear correlation by equation 73. Kaula (1959a, b) applied extensively to 1° by 1° square means the method of Markov estimation in terms of gravity and topography, which estimates for the $(i+1)$th member of a series of step-by-step extrapolations

$$E\{\Delta g_{i+1} \mid h_{i+1}, \Delta g_i, h_i\} = \frac{\int_{-\infty}^{\infty} x P\{\Delta g_{i+1} = x, h_{i+1} \mid h_i, \Delta g_i\} \, dx}{P\{h_{i+1} \mid h_i, \Delta g_i\}} \tag{79}$$

As Baussus (1960) pointed out, this Markov estimation assumes that the covariance function $K(d)$ can be represented as a negative exponential of d. If $K(d)$ is better represented by Hirvonen's form (equation 74), appreciable improvement is obtainable when we use suitably spaced "next-to-nearest" as well as nearest neighbors in estimation (Moritz, 1962b).

The statistics of other functions can be derived as linear transformations of those for gravity anomalies, either with or without the intermediary of a spectral representation: anomalies at higher altitude (Kaula, 1959a; Hirvonen, 1962; Moritz, 1962a, c), geoid heights (Kaula, 1959a), and deflections of the vertical (Kaula, 1959a; Kaula and Fischer, 1959).

Worldwide Treatment

The statistical analysis of local variations described above has been extended to larger areas by Hirvonen (1956) and Kaula (1959a, b). For covariance estimates, Kaula used, in addition to the local samples, eight regional samples covering an area about 10° by 10° with 56 to 115 members each (see figure 2), plus a single world sample consisting of 569 5° by 5° mean anomalies based on observations in 18% or more of their 1° by 1° squares. The covariances obtained are shown in figures 2 and 3. Table 1 gives the spherical harmonic degree variances $\sigma_n^2\{\Delta g\}$ derived from the covariances by spectral analysis, as described by (32) and (29). These results are certainly of the correct order of magnitude, but the analysis needs to be redone with a properly randomized sample (cf., Shapiro and Silverman, 1960). Hirvonen (1956) did not have the computer facilities necessary to make an extensive estimate of covariance, and so he used mainly mean square mean G_s^2 and error of representation E_s^2, based on five regions averaging 8° by 8° plus the data of Tanni (1948). His results imply a more gently varying field than Kaula's.

These authors and Cook (1950, 1951), Kaula (1957), and Molodenskiy et al. (1960) devote considerable attention to estimating the accuracy of determination of the geoid height and slope at specific points, given either the actual data distribution or a hypothetical dense net within a given distance of the point. Hirvonen's and Kaula's studies yield uncertainties that are probably too low, owing to the samples used and the assumption of randomness of 30° by 30° square

TABLE 1.—*Degree Variances of Free Air Gravity Anomalies (Kaula, 1959b)*

n	σ_n^2, mgal^2	n	σ_n^2, mgal^2	n	σ_n^2, mgal^2	n	σ_n^2, mgal^2
2	7	9	22	17	12	25	9
3	44	10	15	18	19	26	11
4	30	11	18	19	10	27	4
5	10	12	7	20	7	28	8
6	24	13	15	21	14	29	5
7	3	14	23	22	10	30	−2
8	23	15	22	23	9	31	1
		16	6	24	11	32	2

means; Cook's and Molodenskiy's studies yield uncertainties that are probably too high because of use of the least-squares-determined spherical harmonic coefficients of Jeffreys (1943) and Zhonglovich (1952), respectively.

The principal point of difference in various attempts to determine the external field from the incomplete information shown in figure 1 has been on whether or not, and how, to utilize the topography as a means of interpolating and extrapolating over great distances. Dubovskiy (Molodenskiy, 1945), and Tanni (1948) used the topography by assuming zero isostatic anomaly in the unsurveyed areas. Kaula (1959b), using the topography by applying the Markov extrapolation technique of equation 79 to 5° by 5° means worldwide, probably obtained results similar to isostatic assumption although less susceptible to bias because of its broader statistical basis. In applying numerical integration, Heiskanen (1957) assumed zero free air anomaly with respect to the

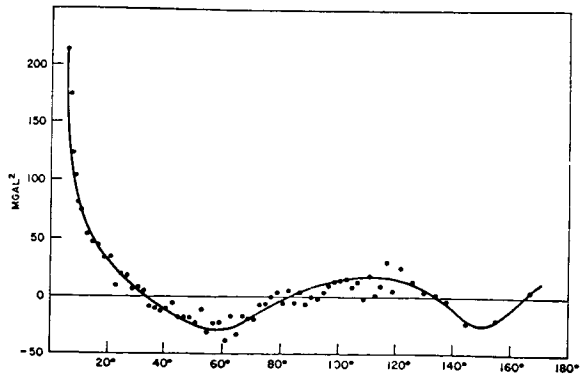


FIGURE 3.—Estimated worldwide covariance of gravity anomalies.

International Formula and Zhonglovitch (1952) assumed zero anomaly with respect to his least-squares-determined third degree figure. Analyses limited to best fits to the observed gravimetry are those of Jeffreys (1943), Zhonglovich (1952) for harmonic degrees two, three, and four, and Uotila (1962). All these solutions employed simple least squares; i.e., the harmonic coefficients were obtained as parameters **z** in (39) with coefficients **C** and covariance matrix **W** of the diagonal form specified by (44). To overcome the ill-conditioning due to neglect of covariance, Jeffreys (1943) grouped his data heavily into 30° by 30° square means and omitted all harmonic coefficients that the normal equation diagonal coefficient and constant indicated as making a small contribution before determining the remaining coefficients through degree 3 by least squares. Zhonglovich (1952) used the 10° by 10° square means and tried several-least squares solutions for different harmonics up through degree 4 before choosing as best a step-wise solution in which the coefficients for each successive degree 2, 3, 4 were determined from the residuals with respect to previously determined degrees. Uotila (1962), benefiting from considerably more data, determined harmonic coefficients up through degree 4 by direct least-squares fit to 5° by 5° mean anomalies, holding the harmonics $(n,m) = (1,0)$, $(1,1)$, and $(2,1)$ zero and $(2,0)$, $(3,0)$, $(4,0)$ fixed at satellite-determined values.

Jeffreys (1959, 1961) rejects use of the topography for long-range estimation; it is not clear whether he does so because the earlier determinations of gravity at sea obtained mostly positive anomalies, because the long-wave correlation is

small [i.e., $\sigma_n(fh)$ small for small n in equation 64], because the applications of the topography have involved the questionable isostatic or Markovian assumptions, or because, as a geophysicist, he is interested in estimating the amplitude, rather than the phase, of the variations. In any case it seems clear that (1) a solution optimum in the sense of quadratic sum minimization will be the one using the most information [i.e., the estimate by (72) will be better than that by (35)]; (2) the optimum estimates will be smaller in the mean square than the true values as a consequence of the smoothing effect of any prediction procedure; and (3) the appropriate measures of amplitude alone are the degree variances σ_n^2.

It is difficult to characterize any of the existing solutions as approximations to the solutions described by (35) and (66), so that it is not clear to what extent they fall short of fully exploiting the available data. In principle, the general solution up to about 12th degree harmonics should be easy with modern computing facilities: the half wavelength for the 12th degree is 15°, there are about 160 15° by 15° squares with observations, and the 160 by 160 covariance matrix can be stored all at once in the core and inverted in a couple of minutes in an IBM 7094 computer. The topography can also be incorporated in a more effective manner than before by using a new development thereof in spherical harmonics by Bruins, (Vening–Meinesz, 1959).

USE OF SATELLITE ORBITS

Observing System

The principal types of satellite tracking that have produced results useful for geodesy are essentially those described in the review of Kaula (1962a): the 500-mm focal length $f/1$ Baker–Nunn tracking cameras; the 1000-mm focal length, $f/5$, modified aerial reconnaissance cameras, fixed and equatorially mounted; the Minitrack 108- and 136-Mc/s radio interferometers; and the Transit 324- to 162-Mc/s radio Doppler trackers. The most important recent instrumental development has been the Anna geodetic satellite (Macomber, 1963), which incorporates magnetically oriented xenon-gas-discharge lamps capable of generating a flash of 8800 candle seconds along the axis, in addition to radio transponders and beacons.

The distribution of the tracking stations of some of the principal systems used for geodesy is shown in figure 4. The nonuniformity of distribution

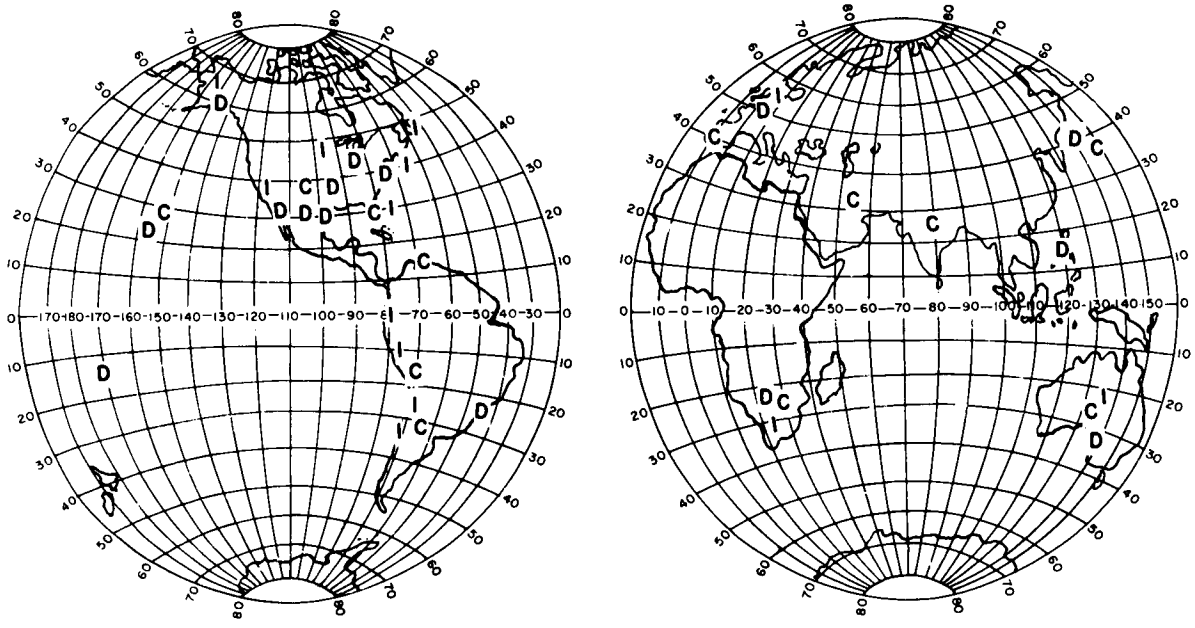


FIGURE 4.—Some satellite tracking systems used for geodesy: C, Baker-Nunn cameras; D, transit Doppler receivers; I, Minitrack radio interferometers and Mots cameras.

gives rise to some statistical problems because of the limited coverage of an orbit by a limited number of stations. Solving the spherical triangle formed by the orbital plane, the equator, and the meridian of a particular longitude λ, we obtain

$$\omega+f=\sin^{-1}(\sin\phi/\sin i) \tag{80}$$

$$\Omega-\theta=\lambda-\cot^{-1}[\pm(\tan^2 i\,\csc^2\phi-\sec^2 i)^{1/2}] \tag{81}$$

where i, ω, f, and Ω are, respectively, the inclination, argument of perigee, true anomaly, and longitude of the node of the orbital ellipse, and θ is the Greenwich sidereal time. Since the inclination i varies but slightly, a station of position (ϕ, λ) can observe the satellite only near certain values of $\omega+f$ and $\Omega-\theta$. For fairly well-distributed networks of about twelve stations, such as the Smithsonian Astrophysical Observatory Baker–Nunn cameras and the Applied Physics Laboratory transit Doppler stations, the distortion due to this effect is probably slight except for series of camera observations dependent on solar illumination over durations short compared to a full cycle of nodal motion with respect to the sun; i.e., less than a few months.

For the purpose of determining orbital variations, the random errors are quite satisfactorily small for all the systems mentioned above except the Minitrack interferometers, for which the effect of ionospheric refraction irregularities on observed directions are of the order of ± 0.0005 to ± 0.001, or $\pm 100''$ to $\pm 200''$. Ionospheric refraction is also probably a source of some systematic error in the Minitrack system. In the transit Doppler system, ionospheric refraction is believed to be largely eliminated by simultaneous transmission on two frequencies, and the most probable source of systematic error is shifting of the reference frequency provided by the oscillator in the satellite, to the extent that the frequency is often taken as a separate unknown for each observed pass. The principal likely systematic error in camera observations is in the timing of observations of sunlit satellites, caused by difficulties of synchronization of the camera shutter with the clock and by error in the propagation delay time, possibly of the order of 0.001 second.

From the point of view of determining the gravitational field, the greatest systematic error common to all observational systems is error in tracking station positions. Error of positions of stations within the same triangulation system with respect to each other should be less than 20 or 30 m; the principal possible exceptions are the positions of stations in South America and South Africa with respect to those in the same system in the northern hemisphere. Errors of position of triangulation systems with respect to each other and to the center of mass of the earth should be less than 100 m for continental systems and less than 500 m for isolated stations. In addition to these errors that should exist, there have been cases where mistakes of as much as a kilometer have existed in station positions.

Dynamics

The aspects of celestial mechanics important to understanding of close satellite orbits are explained in recent texts such as those of Baker and Makemson (1960) and Brouwer and Clemence (1961). Satellite orbit dynamics with emphasis on geodetic applications is discussed by Kaula (1962) and Mueller (1963).

The theoretical problem of satellite orbits is to solve the equations of motion:

$$\ddot{\mathbf{r}}=\nabla(V+R_s)+\nabla_v R_d \tag{82}$$

where ∇ is the gradient with respect to position, ∇_v is the gradient with respect to velocity, V is the earth's gravitational potential as given by (9), R_s is the gravitational plus radiation pressure potentials of third bodies (sun, moon, etc.), and R_d is the atmospheric drag potential, a function of both position and velocity of the satellite. The information we wish to extract from satellite orbits rests in V, and so for the moment we neglect the other two potentials. The three second-order equations (82) are generally reduced to six first-order equations by change of variables:

$$\dot{s}_i=\sum_j C_{ij}(\mathbf{s})\,(\partial V/\partial s_i) \tag{83}$$

The simplest set of six variables $\mathbf{s}$ would, of course, be the position vector components $\{x, y, z\}$ and the velocity vector components $\{\dot{x}, \dot{y}, \dot{z}\}$, referred to inertial space with origin at the earth's center. These six variables can be transformed to the six parameters of a Kepler ellipse with one focus at the origin: $\{a, e, i, \Omega, \omega, f\}$. The relationships be-

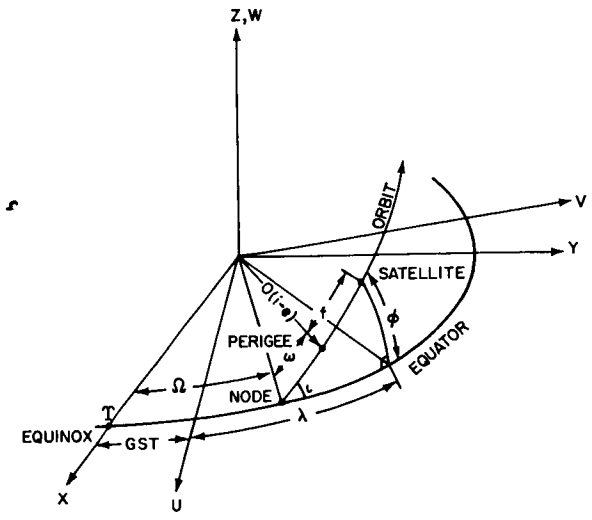


FIGURE 5.—Orbit and coordinate systems.

tween these parameters and the earth-fixed coordinates $\{u, v, w\}$ are shown in figure 5. In the angle $\omega+f$, ω is the argument of perigee, the angle from the ascending node Ω to perigee, the point of closest approach of the ellipse to the origin, and f is the true anomaly, the angle from perigee to the satellite. Alternate ways of expressing the anomaly of the satellite are Kaula [(1962) p. 194] the eccentric anomaly E:

$$\tan\tfrac{1}{2}E=[(1-e)/(1+e)]^{1/2}\tan\tfrac{1}{2}f \tag{84}$$

and the mean anomaly M:

$$M=E-e\sin E \tag{85}$$

The form of (83) in which **s** is the six Keplerian elements is advantageous because in the case of a purely central field; i.e., V is kM/r, the only nonzero rate of change is that of the anomaly; furthermore, for the mean anomaly M it is constant (Kepler's third law):

$$n=\dot{M}=(kM)^{1/2}a^{-3/2} \tag{86}$$

where M on the right is the mass of the central body.

In the case of a satellite moving in the actual potential field of the earth, by (9), for which the departures from a central field are $0(10^{-3})$, the Keplerian elements are still convenient because the rates of variation $\dot{s}_i$ in (83) will be small except for $\dot{M}$, and hence representable as perturbations of the elliptic motion. The form of (83) for Keplerian elements is (Brouwer and Clemence, 1961, p. 289; Kaula 1962, p. 198):

$$\dot{a}=(2/na)\cdot(\partial V/\partial M)$$

$$\dot{e}=\frac{1-e^2}{na^2e}\cdot\frac{\partial V}{\partial M}-\frac{(1-e^2)^{1/2}}{na^2e}\cdot\frac{\partial V}{\partial\omega}$$

$$\omega=-\frac{\cos i}{na^2(1-e^2)^{1/2}\sin i}\frac{\partial V}{\partial i}+\frac{(1-e^2)^{1/2}}{na^2e}\frac{\partial V}{\partial e}$$

$$di/dt=\frac{\cos i}{na^2(1-e^2)^{1/2}\sin i}(\partial V/\partial\omega)-[na^2(1-e^2)^{1/2}\sin i]^{-1}(\partial V/\partial\Omega)$$

$$\dot{\Omega}=[na^2(1-e^2)^{1/2}\sin i]^{-1}\cdot(\partial V/\partial i)$$

$$\dot{M}=n-\frac{1-e^2}{na^2e}\frac{\partial V}{\partial e}-\frac{2}{na}\frac{\partial V}{\partial a} \tag{87}$$

As is indicated by figure 5, it is a purely geometrical problem to transform $V(r, \phi, \lambda)$ as given by (9) to a form $V(\Omega, i, \omega, a, e, M)$ suitable for taking derivatives in (83). The solution which is given by Kaula [1961a]:

$$V=kM\Big[r^{-1}+a^{-1}\sum_{n=2}^{\infty}(a_e/a)^n\sum_{m=0}^{n}N_{nm}\sum_{p=0}^{n}F_{nmp}(i)\cdot\sum_{q=-\infty}^{\infty}G_{npq}(e)\,S_{nmpq}(\omega, M, \Omega, \theta)\Big] \tag{88}$$

where the normalization factor

$$N_{nm}=\left[\frac{(2n+1)\,(n-m)\,!\,(2-\delta_{0m})}{(n+m)\,!}\right]^{1/2} \tag{89}$$

$$F_{nmp}(i)=\sum_{t}\frac{(2n-2t)\,!}{t!(n-t)\,!(n-m-2t)\,!\,2^{2n-2t}}\sin^{n-m-2t}i\cdot\sum_{s}\binom{m}{s}\cos^{s}i\sum_{c}\binom{n-m-2t+s}{c}\binom{m-s}{p-t-c}(-1)^{c-k} \tag{90}$$

in which t is summed from 0 to p or k (defined after equation 11), whichever is less; s, from 0 to m; and c, over all values making the two binomial coefficients nonzero;

$$G_{np(2p-n)}(e)=(1-e^2)^{(1/2)-n}\sum_{d=0}^{p'-1}\binom{n-1}{2d+n-2p'}\binom{2d+n-2p'}{d}(\tfrac{1}{2}e)^{2d+n-2p'} \tag{91}$$

where $p'=p$, $p\leq n/2$, $p'=n-p$, $p\leq n/2$, for $q=2p-n$; for $q\neq 2p-n$, more complicated infinite series are required (Kaula, 1961a, equations 24–26); and

$$\begin{aligned}S_{nmpq}(\omega, M, \Omega, \theta)=&\begin{bmatrix}\bar{C}_{nm}\\ -\bar{S}_{nm}\end{bmatrix}_{(n-m)\text{odd}}^{(n-m)\text{even}}\cos\{(n-2)\omega+(n-2p+q)M+m(\Omega-\theta)\}\\ &+\begin{bmatrix}\bar{S}_{nm}\\ \bar{C}_{nm}\end{bmatrix}_{(n-m)\text{odd}}^{(n-m)\text{even}}\sin\{(n-2p)\omega+(n-2p+q)M+m(\Omega-\theta)\}\end{aligned} \tag{92}$$

Other expressions are given by Musen (1960) and Groves (1960). If V as expressed by equations 88–92 is differentiated and the derivatives are placed in (87), the form of S_{nmpq} in (92) indicates that $\dot{a}$, $\dot{e}$, di/dt must be sinusoidal. If we assume that these sinusoidal variations are small enough that a, e, i can be considered as constant on the right of (87), then (92) further indicates that only $\dot{\omega}$, Ω, and $\dot{M}$ have constant terms, and that these terms arise from cases where n is even, m is zero, p is $n/2$, and q is zero. Hence to a linear approximation, (87) can be integrated assuming a, e, i, n constant (except for n in $\dot{M}$, the variation of which is obtainable from equation 86 and $\dot{a}$) and M, ω, Ω secularly changing.

However, since $\bar{C}_{2,0}$ is $0(10^{-3})$ while all the other $\bar{C}_{nm}$ and $\bar{S}_{nm}$ are $0(10^{-6})$, a linear approximation does not suffice for $\bar{C}_{2,0}$: terms with coefficient $\bar{C}_{2,0}^2$ are required. As usual when a problem becomes nonlinear, it becomes appreciably more complicated, and care must be taken in defining the constants of integration. The obvious choices are the values of the Keplerian elements at a particular instant of time, i.e., the osculating elements. The osculating elements are the constants of integration when the problem is solved in a purely numerical manner.

When the problem is developed further for an iterative or analytic solution, however, it is usually found more convenient to define the constants as elements of a fictitious reference orbit or an intermediate orbit. In some theories this intermediary is defined geometrically, as, e.g., representing all constant and secular terms of the actual orbit; in other theories it can be defined dynamically as corresponding to a constant part of the potential V. Other aspects in which theories may differ are the coordinate system; the independent variable; the orbital elements—often a canonical set is used, i.e., one such that the C_{ij} in (83) consists of only one nonzero constant per row; and the point at which the development of the problem changes from algebraic to numerical. This large variety of possibilities gave rise to a large number of theoretical papers in the period 1957–1961 on the orbit of a close satellite of an oblate planet to at least $0(J_2^2)$ in secular motions (J_2 is $-5^{1/2}\bar{C}_{2,0}$). The theories that probably have been used the most are those of Musen (1959), who adapted the Hansen lunar theory to a form suitable for solution by iteration; Brouwer (1959), who applied Von Zeipel's method of canonical transformation with a purely Keplerian intermediary; Kozai (1959a, b), who extended the

Lagrangian equations 87 to higher order terms; Vinti (1959, 1961) and Izsak (1960), who both separated the equations of motion by using ellipsoidal coordinates; King–Hele (1958), who employed a Keplerian ellipse of fixed inclination and perigee argument as intermediary and solved in successive approximations according to powers of J_2 and e; and Merson (1961), who made a development similar to King–Hele's starting from osculating elements when the satellite is at the node. (See Kaula (1962) for more description and comparison.)

Since 1960, more theoretical attention has turned to the problems of resonance associated with close satellites. The problem of critical inclination in the vicinity of $\sec^{-1} 5^{1/2}$, or 63°26′, at which perigee motion is zero, is of relatively little geodetic interest; the most complete solution is probably that of Izsak (1962), as extended by Aoki [1963]. Of much more geodetic significance are the 24-hour orbits (semimajor axis 42,000 km), which resonate with tesseral harmonics $\bar{C}_{nm}$, $\bar{S}_{nm}$ for which $n-m$ is even, since

$$\dot{\Omega}+\dot{\omega}+\dot{M}-\dot{\theta}\approx 0 \qquad (93)$$

The problem of orbits to which (93) applies with potential

$$V=(kM/r)+R$$
$$=kM[r^{-1}+(a_e^2/a^3)N_{22}F_{220}G_{200}S_{2200}] \qquad (94)$$

has been analyzed by Blitzer et al. (1962, 1963), who used a linearized system of differential equations; by Musen and Bailie (1962), who isolate $(\Omega+\omega+M-\theta)$ as a canonical element and develop the Hamiltonian in powers of the ratio $(\partial R/\partial L)/(\partial^2 R/\partial L^2)$, where $L=(kMa)^{1/2}$; and by Morando (1963), who applies Von Zeipel's method and develops the determining function in powers of $(C_{23}^2+S_{22}^2)^{1/4}$.

Elaborately developed theories such as those of Musen (1959) and Brouwer (1959) are advantageous to analyze secular changes for zonal harmonics and to conserve computer time when observations are infrequent (say, an average of less than one per hour). Computation of orbital arcs of a few days or less using frequent radio tracking is still done mostly by numerical integration.

As necessary as a correct dynamical theory, and often more laborious, are differential correction systems to determine from observations the numerical values to use in the theories. Examples of differential correction programs for close satellites are those of Veis and Moore (1960) and Merson (1963). The geometrical aspects of differential correction schemes are fairly complicated but straightforward (see Veis (1960) and Kaula (1961a, 1962) for details). The statistical aspects are not so complicated, but are much less satisfactorily treated, partly because a rigorous treatment would require excessive computer storage and time, and partly because adequate statistical models are lacking. Satellite orbit observations always have an appreciable amount of serial correlation, principally because of the inability to account for air drag. However, the statistical analysis of satellite orbit accelerations has received relatively little attention, the only active worker thereon being Moe (1963). No differential correction method takes into account correlation between observations of different passes, as suggested in the section of this review on time series. The most elaborate treatment existing (Kaula, 1963b) allows for correlation within a pass and applies various other devices such as giving greater weight to the across-track than to the along-track component of an observation; preassigning variances and covariances of parameters being determined in the analysis, so that mathematically they become "observations"; and weighting observations according to their distribution with respect to phase angles believed critical.

Determination of Zonal Harmonics

For the principal secular or long-period effect of a zonal harmonic, $m=0$, the disturbing terms in (88) can be taken as:

$$\text{Long period } R_{no}=-kMJ_n(a_e^n/a^{n+1})$$
$$\cdot F_{nok}(i)G_{nk(2k-n)}(e)\begin{cases}1, & n \text{ even}\\ 2\sin\omega, & n \text{ odd}\end{cases} \qquad (95)$$

where J_n is $-(2n+1)^{1/2}\bar{C}_{n,0}$. The linear perturbations can be obtained by using this disturbing function in the Lagrangian equations of motion 87; in addition, nonlinear terms J_2^2 and J_2J_n, n odd, need to be taken into account.

The customary method of determining the J_n's from satellite orbits is to analyze the long-term variation of orbital elements determined by a differential correction fit to observations over a few days at a time. Precautions that must be taken in analyzing secular changes of the node Ω and perigee ω to determine even degree zonal harmonics include:

1. The set of satellite orbits used should have a variety of inclinations sufficient to separate the different harmonics.

2. The constants of integration determined by analyzing observations must be consistent with the algebraic form of the terms containing J_2^2 ($\bar{C}_{2,0}^2$ in our notation).

3. The mean value of the constants of integration must be accurately determined for the duration from which the secular rates $\dot{\Omega}$, $\dot{\omega}$ are determined. It is also to be emphasized that mean values of the constants of integration will differ from mean values of osculating elements for some theories (e.g., the inclination in the theory of King–Hele (1958), or the mean motion in the theory of Brouwer (1959). Averaging of the elements a and e is important in order to remove secular drag effects; if the a, e used are not average values, but rather values for some epoch noncentral to the duration used, about the best that can be done is to use the perturbation in mean anomaly to correct node and perigee assuming the perigee height to remain fixed (O'Keefe et al., 1959; Kozai, 1962, 1963a):

$$\Delta(\Omega, \omega) = \frac{(\dot{\Omega}, \dot{\omega})}{3n} \frac{7-e}{1+e} \Delta M \qquad (96)$$

The effects of errors Δe, Δi, Δa in the orbital elements on determination of the secular rate of the node are (Kozai, 1962):

$$\Delta\dot{\Omega} = \dot{\Omega}\{[4e/(1-e^2)]\Delta e - \tan i\, \Delta i - (7/2a)\Delta a\} \qquad (97)$$

For some satellites of high inclination, error in determining zonal harmonics will thus come more from error in the mean inclination than in $\dot{\Omega}$ itself.

4. If luni-solar attraction, radiation pressure, and other perturbations are not removed in determining the mean values of the constants of integration, they can distort determination of the rates $\dot{\Omega}$ and $\dot{\omega}$ not only through purely secular effects but also through periodic perturbations. A periodic perturbation $\Delta(\Omega, \omega)\sin\{\kappa t - \lambda\}$ will affect the apparent secular rate from observations lasting from t_1 to t_2 by an amount $\delta(\dot{\Omega}, \dot{\omega})$:

$$\delta(\dot{\Omega}, \dot{\omega}) = \frac{\Delta(\Omega, \omega)[\sin\{\kappa t_2 - \lambda\} - \sin\{\kappa t_1 - \lambda\}]}{t_2 - t_1} \qquad (98)$$

5. If the perturbations are removed in determining the constants of integration, in addition to direct effects $\Delta_1(\Omega, \omega)$, the interaction of perturbations Δe, Δi with the secular effect of J_2 may cause an indirect effect $\Delta_2(\Omega, \omega)$ large enough that it should be taken into account:

$$\Delta_2(\Omega, \omega) = \frac{\partial(\dot{\Omega}, \dot{\omega})}{\partial e} \int \Delta e\, dt + \frac{\partial(\dot{\Omega}, \dot{\omega})}{\partial i} \int \Delta i\, dt \qquad (99)$$

6. Satellites with low perigee, nonspherical shape, and large area-to-mass ratio should be avoided because of the difficulty in calculating drag and radiation pressure effects on the mean elements and the secular rates, particularly on $\dot{\omega}$.

7. Use of orbits provided by routine prediction services should be avoided, because of possible imprecision in the definition of orbital elements and inaccuracy in the determination thereof.

The most important recent analyses of secular motions have been by Kozai (1962) and King–Hele et al. (1963). Kozai uses $\dot{\Omega}$ and $\dot{\omega}$ of thirteen satellites ranging from 32.9° to 66.8° in inclination. However, he weights the data inversely proportional to the squares of the standard deviations, which vary greatly, so that the result is almost entirely determined by the $\dot{\Omega}$ of $1960\iota_2$ (inclination 47.2°) and the $\dot{\Omega}$ and $\dot{\omega}$ of $1959\alpha_1$ and 1959η (inclinations 32.9°, 33.4°). King–Hele et al. (1963) used only the nodal motion $\dot{\Omega}$ of seven satellites at widely spaced intervals of inclination from 32.9° to 97.4° and weighted each satellite about equally. However, to obtain inclinations above 53.8°, they used elements provided by routine prediction services. The disagreement between the results of Kozai (1962) and King–Hele et al. (1963) in table 2 at this late date is disappointing. What are required are accurately determined orbits of high inclination. A limited amount of the requisite data are now available in the form of 552 precisely reduced Baker–Nunn

TABLE 2.—*Zonal Harmonic Coefficients of the Gravitational Field*

Coefficient*	Newton et al. (1961)	Smith (1961, 1963)	Kozai (1962)	Shelkey (1962)	King-Hele et al. (1963)	Anderle and Oesterwinter (1963)
$J_2\times10^6$		1083.15	1082.48	1082.61	1082.86	1082.47
$J_3\times10^6$	−2.36	−2.44	−2.56	−1.94		−2.48
$J_4\times10^6$		−1.4	−1.84	−1.52	−1.03	−1.40
$J_5\times10^6$	−0.19	−0.18	−0.06	−0.41		−0.14
$J_6\times10^6$		0.7	0.39	0.73	0.72	
$J_7\times10^6$	−0.28	0.30	−0.47			
$J_8\times10^6$			−0.02		−0.34	
$J_9\times10^6$			0.12			
$J_{10}\times10^6$					0.50	
$J_{12}\times10^6$					0.44	

*$J_n = -(2n+1\ \bar{C}_{n0})^{\frac{1}{2}}$

camera observations over 54 days of $1961\alpha\delta_1$, which has inclination 95.9°, perigee height 3500 km. The observed nodal motion $\dot{\Omega}$ determined as a byproduct of analysis for tesseral harmonics (Kaula, 1963d) is $+0.21037°\pm0.00010$/day. Including a luni-solar effect of $-0.00006°$/day, the calculated motion $\dot{\Omega}$ is $+0.21034°$/day using the J_2 through J_8 of Kozai (1962), and 0.21056°/day using the J_2 through J_{12} of King-Hele et al. (1963).

The determination of the odd degree zonal harmonics is somewhat easier, since there do not appear to be any other significant effects of period $2\pi/\dot{\omega}$. The principal precautions are to include the

TABLE 3.—*Determinations of GM*

Method	Reference	Sources of error	*GM* 10^{21} cm^3 sec^{-2}
Terrestrial geodesy	Fischer (1962)	Triangulation	3.986040
	Kaula (1961b)	Gravimetry	3.986020±0.000028
	a_e from Kaula (1961*b*)		
	γ_e from Uotila (1962)		
Lunar motion and radar distance.	Yaplee et al. (1963)*	Lunar radius	3.986057
Lunar motion and triangulated distance.	Fischer (1962)*	Lunar radius	
	Crommelin a	Triangulation	3.986451
	Fischer (1962)*		3.986078
	O'Keefe and Anderson a		
Lunar probe and Doppler	Hamilton et al. (1963)	Observational Station position	3.986016±0.000025
Close satellite and Doppler	Anderle and Oesterwinter (1963).	Observational Orbit perturbations	3.985889
Close satellite and camera	Kaula (1963b), 1960_{42}	Triangulation	3.986037±0.000012
	Kaula (1963d), $1961\alpha_{61}$	Orbit perturbations	3.985993±0.000011

*These references are sources for only the distance to the nearest point of the moon; a lunar radius of 1738.7 km and a lunar mass of 1/81.3015 have been used to calculate the values of *GM* given in the last column.

J_2J_n terms, following (99), and the drag correction of (96) (Kozai, 1959a). Results by the principal investigators are given in table 2.

Determination of GM

The rapid motion of artificial satellites suggests that their mean motion may serve as an accurate method of determining GM, the product of the gravitational constant and the earth's mass, through Kepler's equation 86. Drag and radiation pressure have their greatest effects on the mean anomaly, however, and so it is particularly desirable to have a satellite for which these effects are minimized and readily calculable: spherical, low area-to-mass ratio, small eccentricity, high inclination, perigee height above 1000 km. Determinations of GM from such satellites have been made as byproducts of tesseral harmonic analyses by Kaula (1963b, d), using camera observations, and by Anderle and Oesterwinter (1964), using Doppler observations. The determination from camera data depends on the geodetic triangulation connecting tracking stations to establish scale; that from Doppler data may be affected by ionospheric refraction effects. A more distant satellite would reduce these effects: in this category are the use of a lunar probe (Hamilton et al., 1963) and those methods measuring the distance of the moon by radar (Yaplee et al., 1963) and by triangulation (Fischer, 1962). The radar method now obtains an internal accuracy of ± 200 m for the distance of the moon by correcting for the variations in topography on the moon's surface. Both the radar and triangulation methods are affected by error in the radius in the moon in the direction of the earth, for which the scatter of different determinations is about 2 km, equivalent to a variation of 0.00007×10^{14} m^3/sec^2 in GM. Another possible source of error in the ratio μ of the moon's mass to the earth's, appearing in the modified Kepler equation

$$GM = n^2a^3[(1+\beta)^3/(1+\mu)] \qquad (100)$$

where β is the effect of the sun on the mean distance. When the astronomical unit found by radar measurements to Venus is used, the values of μ^{-1} deduced from the lunar inequality (the monthly revolution of the earth about the center of mass of the earth and moon) based on observations of Eros range from 81.26 to 81.36. A new determination of the inequality from Doppler observations of the Mariner 2 Venus probe yields a preliminary value of 81.3015 ± 0.0033 (Hamilton et al., 1963).

Determination of Tesseral Harmonics

The determination of tesseral harmonics from satellite orbits depends principally on the terms in the disturbing function, (88), for which the argument is $\{(n-2p)\omega + m(\Omega-\theta)\}$, and for which $|q|$ is small. Because of the unavoidable inclusion of the earth's rotation rate $\dot{\theta}$, the frequencies involved differ by orders of magnitude from those involved in analysis for zonal harmonics, and so the problems are quite different. On the one hand, long-term-drag, luni-solar, and radiation pressure effects are of little influence; on the other hand, errors in station position and nonuniform distribution of observations become important. Usually the frequency of observation is not much higher than the frequencies of the orbital perturbations caused by the tesseral harmonics; hence it is not possible to employ the method of first determining smoothed osculating elements from the observations and then analyzing variations in these elements. The analysis must be made directly in terms of the observations themselves; i.e., there must be formed a partial derivative of each observation with respect to each of the parameters sought. A rigorous solution will then follow that of the times series of equations 57–60, in which $x_f(t)$ are the Keplerian elements, p_l are the gravitational coefficients, y_j are the observations, and q_m are the corrections to station position coordinates. The difficulty in the rigorous solution, as stated, is taking into account the covariance between observations, since this entails arrays of dimension comparable to the number of observations. Furthermore, the statistical analysis of drag effects has not been developed sufficiently (Kaula, 1961c; Moe, 1963). Consequently, in practice all analyses have neglected covariance between observations of different passes and have applied various devices to minimize the effect of this neglect:

1. Higher weighting of the across-track than of the along-track component of an observation is used, because a proportionately much greater

TABLE 4.—*Harmonic Coefficients of the Gravitational Field*

Coefficient	Kaula (1961b)	Uotila (1962)	Kozai (1963b)	Izsak (1963)	Anderle and Oesterwinter (1964)	Kaula (1963d)	Guier (1963)
$\bar{C}_{20}\times10^6$	−484.23	(−484.10)	(−484.10)	(−484.10)	−484.09	−484.08	
$\bar{C}_{22}\times10^6$	0.75	0.69	1.11	1.50	2.85	1.88	2.60
$\bar{S}_{22}\times10^6$	−0.38	−2.25	−1.47	−0.62	−1.53	−1.38	−0.99
$\bar{C}_{30}\times10^6$	0.78	(0.91)	(0.97)	(0.97)	0.94	0.97	----------
$\bar{C}_{31}\times10^6$	1.03	0.10	1.75	1.04	----------	1.52	1.64
$\bar{S}_{31}\times10^6$	0.39	−0.63	0.14	0.06	----------	0.14	0.18
$\bar{C}_{32}\times10^6$	0.97	1.19	0.34	0.27	----------	−0.02	0.84
$\bar{S}_{32}\times10^6$	−0.03	−0.22	−0.24	−0.53	----------	0.42	−0.07
$\bar{C}_{33}\times10^6$	0.57	1.50	−0.45	0.51	----------	0.70	1.06
$\bar{S}_{33}\times10^6$	1.40	2.44	0.60	0.89	----------	0.76	1.01
$\bar{C}_{40}\times10^6$	0.25	(0.57)	(0.61)	(0.61)	0.47	0.67	----------
$\bar{C}_{41}\times10^6$	−0.63	−0.15	−0.30	−0.30	−0.71	−0.33	−0.60
$\bar{S}_{41}\times10^6$	−0.15	−0.20	−0.46	−0.34	−0.40	0.37	−0.49
$\bar{C}_{42}\times10^6$	0.46	0.82	−0.18	0.16	0.45	0.01	0.27
$\bar{S}_{42}\times10^6$	0.42	0.46	0.21	0.55	1.20	0.35	1.19
$\bar{C}_{43}\times10^6$	0.51	1.20	0.59	0.36	2.64	0.17	1.33
$\bar{S}_{43}\times10^6$	−0.01	−0.63	0.02	0.25	−0.60	0.41	−0.05
$\bar{C}_{44}\times10^6$	−0.10	0.66	0.78	0.46	----------	−0.01	−0.37
$\bar{S}_{44}\times10^6$	0.36	−0.12	1.26	0.77	----------	0.18	0.31

part of the drag effect is in the mean anomaly than is true for the gravitational effects (Kaula, 1963b, Izsak, 1962).

2. The duration for which a set of reference elements, or constants of integration, are determined is limited to one to four weeks. This measure is also desirable to keep the residuals down to not more than a small multiple of the anticipated gravitational effects.

3. Arbitrary polynomials are used to represent some of the variation in mean anomaly.

4. The observations are weighted inversely to their density with respect to the angle $(\Omega-\theta)$ in an attempt to restore some of the orthogonality of gravitationally caused variations to the variations caused by drag (Kaula, 1963b).

The nonuniform distribution of observations arises from the geometrical effect described by (75) plus, in the case of camera observations, from dependence on clear weather and solar illumination of the satellites. This nonuniform distribution, the large number of parameters involved, and the similarity of the perturbations by different gravitational terms (l, m) and (n, m) for which $(l-n)$ is even, all combined to produce an ill-conditioned least-squares solution from observations of a single satellite. The desirable remedy is to combine several arcs of several satellites of widely varying inclination in one solution. However, such a solution would involve well over 100 unknowns, so further compromises must be made. These compromises have been of two principal types:

1. Stepwise least squares. This method is employed by Izsak (1963), Anderle and Oesterwinter (1964), and others. For camera observations there are two steps: in the first step, the reference elements for each orbital arc are determined holding fixed the gravitational coefficients and station positions; in the second step, the orbital elements are held fixed, while the gravitational coefficients and station positions are determined from the residuals of the observations with respect to these reference orbits. For Doppler observations, there are three steps: the above two steps are preceded by a step in which frequency constant and frequency drift corrections are determined for each pass. The likely defect of these methods is that the parameters determined in the earlier steps will absorb some of the effects of the

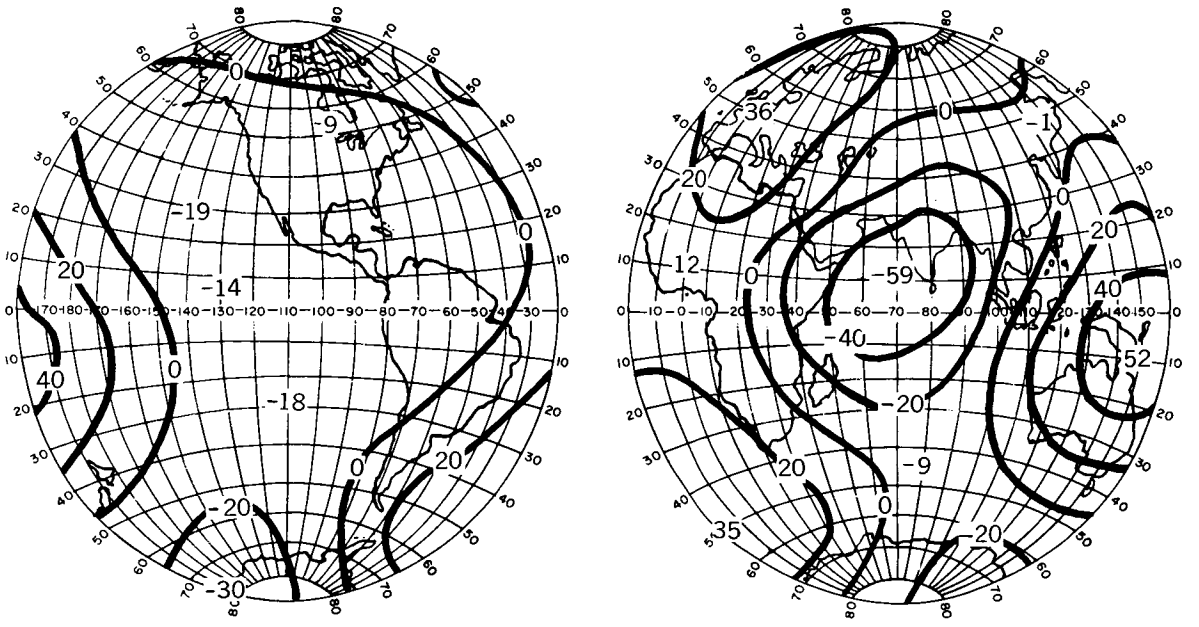


FIGURE 6.—Geoid heights referred to an ellipsoid of flattening 1/298.24, based on camera observations of satellites $1959\alpha_1$, 1959η, $1960\iota_2$, $1961\delta_1$, $1961\alpha\delta_1$, (Kaula, 1963d).

gravitational coefficients and station position shifts. This defect may be particularly severe in the case of orbital arcs with relatively few camera observations. Izsak (1963) omits the worst of such arcs, but this measure introduces further chance of bias because the arcs so omitted will tend to be those for which the apogee is in the southern hemisphere, away from most of the stations. Both Izsak (1963) and Anderle and Oesterwinter (1964) make solutions in which all station positions are allowed to move freely with respect to each other; however, computer program limitations do not permit them to determine all the gravitational coefficients compatible with the accuracy implied by this method.

2. Preassigned covariance matrix (Kaula, 1963b, d). For each orbital arc, the gravitational coefficients and station positions are determined at the same time as the reference elements. To keep the solution from "blowing up" because of ill-conditioning, a covariance matrix $\mathbf{W}_x$ is preassigned to the gravity and position parameters in accordance with (50). The variances in $\mathbf{W}_x$ for the gravitational coefficients are based on the degree variances in table 1; those for the positions of major datums are based on the results of the world geodetic system solution of Kaula (1961b), and those for the positions of isolated stations on statistical analyses of deflections of the vertical. Stations connected by geodetic triangulation are assumed to translate together, consistent with an accuracy of about ± 20 m. The principal defect of the preassigned variance and covariance method is that it may influence the relative magnitude of the results, particularly for harmonics causing similar frequency perturbations in the orbit.

Earlier estimates of tesseral harmonics from satellite orbits (Izsak, 1961a); (Kaula, 1961c); (Kozai, 1961); Newton, 1962) yielded a wide scatter of results and are of interest now only for some discussion of methods. Table 4 gives the most recent results of most of the principal investigators; those by Kozai (1963b), Izsak (1963), and Kaula (1963d) are based on camera data; those by Anderle and Oesterwinter (1964) and Guier (1963) are based on Doppler data. In addition to the gravitational coefficients given in the table, Kaula (1963d) determined 10 tesseral coefficients of degrees 5 and 6, plus 18 datum coordinate shifts; Anderle and Oesterwinter (1964) also determined 54 station coordinate shifts. The most recent determinations from terrestrial data (Kaula, 1961b; Uotila, 1962) are also given in

table 4. Figure 6 is the geoid map corresponding to the solution by Kaula (1963d).

USE OF ASTROGEODESY

The oldest sources of information about the variations in the gravitational field are the slopes of the geoid with respect to the surface of a reference ellipsoid determined from the differences between the astronomic and geodetic positions:

$$\xi = \phi_a - \phi_g$$
$$\eta = (\lambda_a - \lambda_g)\cos\phi \qquad (101)$$

Along a line of geodetic triangulation or traverse in azimuth A the slope will be:

$$t = \eta \sin A + \xi \cos A \qquad (102)$$

and for the change in geoid height over an arc distance Ψ there is conventionally applied the simple integration

$$N_B = N_A - R\int_0^{\psi} t\, dl \qquad (103)$$

where R is the radius of the earth.

However, if the triangulation from which the geodetic positions (ϕ_g, λ_g) were calculated with a scale correction $R/(R+h)$ for reduction to sea level altitude h (the "development method"), then an error will accumulate because of neglect of the further scale correction $R/(R+N)$ for reduction to ellipsoid (the "projection method"). After a few thousand kilometers, this error will make a difference of several tens of meters in N_B calculated by (103). The elimination of this discrepancy is known as the Molodenskiy correction; given t, N calculated by (102) and (103) from development-computed ϕ_g, λ_g ,the correct slope θ and height ζ will be (Molodenskiy et al., 1960, p. 33)

$$\theta_B = t_B + (\theta_A - t_A)\cos\psi + (\zeta_A - N_A)\sin\psi + \frac{1}{R}\int_0^{\psi} N\cos(l-\psi)\,dl$$
$$\zeta_B = N_B + R(t_A - \theta_A)\sin\psi + (\zeta_A - N_A)\cos\psi + \int_0^{\psi} N\sin(l-\psi)\,dl \qquad (104)$$

If geodetic control computed by the development method is adjusted forcing conditions on large circuits properly applicable only to pro-

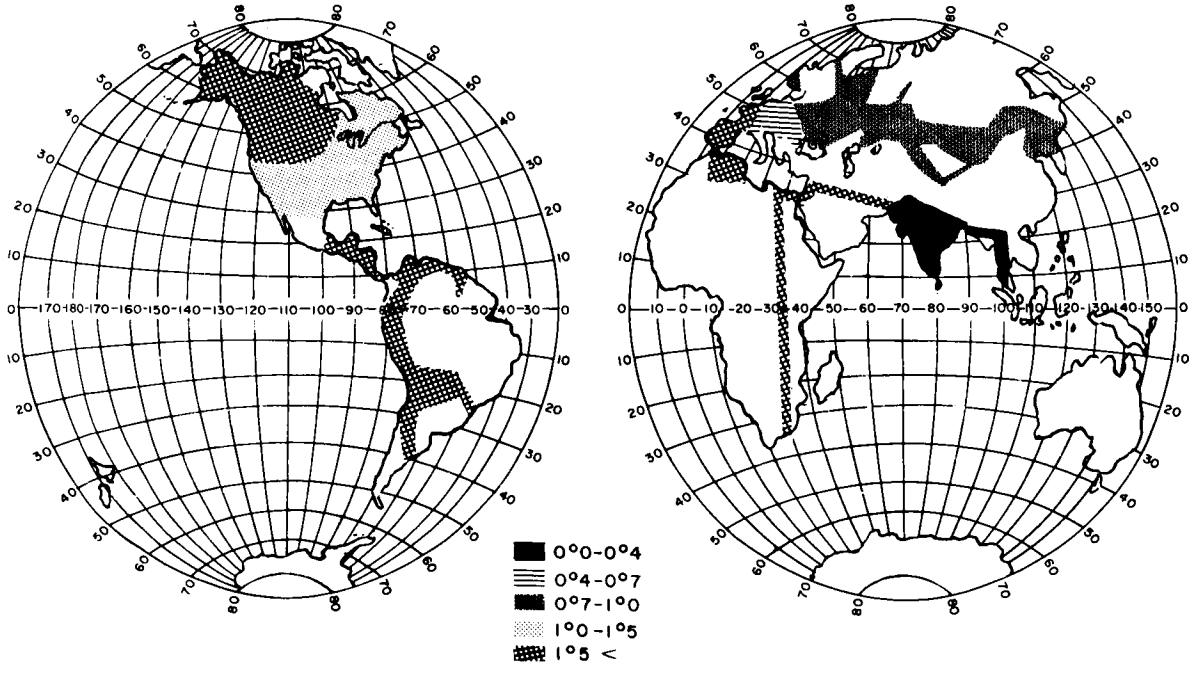


FIGURE 7.—Distribution of astrogeodetic observations.

jection-computed control, then (104) will not remove all error. However, circuits large enough for this distortion to be significant are rare.

The most extensive application of (101) through (104) to astrogeodetic data has been by Fischer (1959a, b; 1960a, b; 1961). Figure 7 shows the distribution of the data she has applied. The limited geographic extent of the astrogeodetic data makes it primarily of value as an independent check on the results obtained by gravimetric and satellite means. The significance of the various spacings of the astronomic stations in figure 7 is that the principal source of error in determining astrogeodetic geoid height differences is the error in determining the continuously varying geoid slope by interpolation between the observed point values. For station spacings S in kilometers, the mean square expected error in the geoid height $E_s\{\epsilon_i^2\}$ derived from the degree variances in table 1 (extended to higher degrees) is approximated by (Kaula, 1961b):

$$E_s\{\epsilon_i^2\} \approx (0.019S - 1.4)^2 \qquad (105)$$

This rule probably gives an underestimate for the geoid height error resulting from arcs carried through rugged topography such as that down the west coast of South America. However, it seems safe to conclude that the rms error for the relative location with respect to each other of points in the major systems shown in Figure 7 is of the order of ± 15 m in the radial coordinate, as well as the two horizontal. This estimate is confirmed by misclosures of less than 25 m for 10,000-km loops around the Caribbean (Fischer, 1959) and around the Black and Caspian seas and through Turkestan (Fischer, 1961). (See Bomford (1960; 1962, pp. 143–159, 325–327) for more detailed discussion of triangulation and geoid height accuracy, and Rice (1962) for an example of an optimum geoidal section survey.)

There are several ways of combining astrogeodetic and gravimetric geoid data. In the U.S.S.R., gravimetry is used to interpolate deflections between astrogeodetic stations (Molodenskiy et al., 1960, pp. 125–141). The traditionally advocated method of combination is to compare the astrogeodetic and gravimetric geoid heights at a few carefully selected points and then to minimize the discrepancy between the two by adjusting the datum position and ellipsoid parameters (Heiskanen and Vening–Meinesz, 1958 pp. 299–310). This method has been applied extensively by Rice (1952) to sixteen stations in the United States and by Szabo (1962) to 6 stations in the United States and 23 stations in Eurasia. However, since there is still appreciable error in the gravimetrically computed deflection at even the best points (probably at least $\pm 1.5''$) the logical conclusion is to compare astrogeodetic and gravimetric geoids wherever the former is available. Zhonglovich (1956) made such a comparison by minimizing the sum $\sum\{(\eta_a-\eta_g)^2+(\xi_a-\xi_g)^2\}$ for the mean values of ninety-six 4° by 4° squares covering North America, using his geoid (Zhonglovich, 1952) for η_g, ξ_g. The mean curvature of the ellipsoid he believed best fitting yields an equatorial radius of $6{,}378{,}104 \pm 42$ m when the flattening of 1/298.3 is enforced. Fischer (1960, 1961) minimized the sum $\sum(N_a-N_g)^2$ at 301 points at 5° intervals throughout the area pictured in figure 7, using N_g calculated by Kaula (1959a, b). Fixing the flattening 1/298.3, the equatorial radius obtained varied from 6,378,160 to 6,378,166 m on different assumptions. Kaula (1961b) used the same data as Fischer, but applied a much more detailed statistical treatment, in which 106 condition equations were written requiring that the corrected differences in astrogeodetic geoid height between mean values for 10° by 10° squares agree with the difference calculated from the corrected harmonic coefficients of the gravitational field. The quadratic sum minimized was

$$\text{minimum} = \mathbf{x}_A\mathbf{W}_A^{-1}\mathbf{x}_A + \mathbf{x}_G\mathbf{W}_G^{-1}\mathbf{x}_G + \mathbf{x}_s\mathbf{W}_s^{-1}\mathbf{x}_s \qquad (106)$$

where $\mathbf{x}_A$ comprises the corrections to the 106 astrogeodetic geoid height differences; $\mathbf{W}_A$ is the associated covariance matrix, taking into account interpolation error, error of representation, and the discrepancy between the actual geoid and that represented by harmonics to the 8th degree; $\mathbf{x}_G$ comprises the 81 spherical harmonic coefficients through the 8th degree; $\mathbf{W}_G$ is the covariance matrix produced by the analysis of gravimetry by Kaula (1959a, b); and $\mathbf{x}_s$ and $\mathbf{W}_s$ pertain to supplementary measurements of secular and long periodic satellite motions and geoid

height matching between adjacent but unconnected datums. A measure of the agreement between astrogeodetic and gravimetric geoids was the value obtained for the quadratic sum of (106); it was 44% higher than the mean χ square expectancy. Increasing the standard deviations accordingly, 6,367,163±15 m and 1/298.24±0.01 were obtained for the ellipsoid parameters. The values of the harmonic coefficients obtained through the 4th degree are given in table 4. The standard deviations obtained for the tesseral harmonic coefficients of degree n averaged about $\pm 0.9\times10^{-6}/(n-1)$. The rms discrepancies from the satellite solution of Kaula (1963d) are $\pm 1.0\times10^{-6}$ for degree 2, $\pm 0.6\times10^{-6}$ for degree 3, and $\pm 0.3\times10^{-6}$ for degree 4, so that the disagreement is one which can be reasonably expected.

CONCLUSIONS

The application of more elaborate statistical techniques made possible by modern computers could very probably extract more information about the gravity field from existing data, both gravimetric and satellite, and would facilitate the planning of programs of additional observations. However, these elaborations would be added to what is already a rather complicated task of measurement, data processing, and theoretical analysis, such that both the benefits and penalties of any modifications are difficult to predict. Also, this review has not touched on the geophysical application of gravity field. It can be argued that an appreciable effort to determine the gravity field is not worth while because knowledge is lacking of the long-term rheology of the earth's interior as well as the mathematical techniques to apply all but the simplest rheologies. Certainly much of what has been discussed in this review is not necessary for the application of gravimetry to study local and regional variations in the crust. What part of it is useful to study variations on a continental or oceanic scale and the implications thereof as to the state of the mantle is perhaps worthy of consideration: To what harmonic degree should the field be developed to apply to ideas of a weak upper mantle or to convection currents? Are the degree variances a useful tool? Should coefficients be derived by techniques which yield mean square values smaller than those of the actual coefficients? The answer is probably the customary one that we cannot predict what the future will want, and hence should strive for the most accurate representation possible.

ACKNOWLEDGMENTS

I am grateful to Mrs. Irene Fischer, Mr. D. G. King-Hele, Dr. Yoshihide Kozai, Dr. M. Caputo, and Dr. H. Moritz for comments and corrections on the manuscript.

REFERENCES

ALLAN, T. D., DEHLINGER, P., GANTAR, C., MORELLI, C., PISANI, M., and HARRISON, J. C., Comparison of Graf-Askania and LaCoste-Romberg surface ship gravimeters, *J. Geophys. Res.*, **67**, 5157–5162, 1962.

ANDERLE, R. J., and OESTERWINTER, C., A preliminary potential for the earth from Doppler observations of satellites, in *Space Research, Proc. Intern. Space Sci. Symp., 4th, Warsaw*, 1963, North-Holland Publishing Co., Amsterdam, to be published, 1964.

AOKI, S., Contribution to the theory of critical inclination of close earth satellites, *Astron. J.*, **68**, 355–358, 1963.

ARLEY, N., and BUCH, K. R., *Probability and Statistics*, John Wiley & Sons, New York, 236 pp., 1950.

ARNOLD, K., Zur Bestimmung der Geoidundulationen aus Freiluftanomalien—Schwerewerte in grossen Höhen über der Erdoberfläche, *Veroeff. Geod. Inst. Potsdam*, **12**, 69 pp., 1959a.

———, Strenge Theorie der absoluten Lotabweichungen als Funktionen der Freiluftanomalien der Schwere, *Veroeff. Geod. Inst. Potsdam*, **13**, 46 pp., 1959b.

———, Numerische Beispiele zur strengen Theorie der Figur der Erde, *Veroeff. Geod. Beitr. Geophys.*, **69**, 140–149, 1960a.

———, Numerische Beispiele zur strengen Theorie der Figur der Erde, *Veroeff. Geod. Inst. Potsdam*, **16**, 66 pp., 1960b.

BAKER, R. M. L., Jr., and MAKEMSON, M. W., *An Introduction to Astrodynamics*, Academic Press, New York, 358 pp., 1960.

BARTLETT, M. S., *An Introduction to Stochastic Processes*, Cambridge University Press, New York, 312 pp., 1956.

BAUSSUS, H. G., A unified isostatic and statistical theory of gravity anomalies and its significance, *Pres. Intern. Assoc. Geodesy, 12th Assembly IUGG*, Helsinki, 33 pp., 1960.

———, Some new aspects regarding the estimation of field variables in meteorology and physical geodesy, *Proc. Symp. Geodesy in Space Age, Publ. Inst. Geod. Phot. Cart.*, **15**, 140–147, 1961.

BJERHAMMAR, A., A general method for an explicit determination of the shape of the earth from gravimetric data, *Bull. Geod.*, **65**, 215–220, 1962a.

———, *On an Explicit Solution of the Gravimetric Boundary Value Problem for an Ellipsoidal Surface of Reference*, Royal Institute of Technology, Stockholm, 93 pp., 1962b.

BLACKMAN, R. B., and TUKEY, J. W., *The Measurement of Power Spectra*, Dover Publications, New York, 190 pp., 1959.

BLITZER, L., BOUGHTON, E. M., KANG, G., and PAGE, R. M., Effect of ellipticity of the equator on 24-hour nearby circular satellite orbits, *J. Geophys. Res.*, **67,** 329–335, 1962.

———, KANG, G., and MCGUIRE, J. B., The perturbed motion of 24-hour satellites due to equatorial ellipticity, *J. Geophys. Res.*, **68,** 950–952, 1963.

BOCHNER, S., *Harmonic Analysis and the Theory of Probability*, University of California Press, Berkeley, 176 pp., 1955.

BOMFORD, G., The junction of the Indian and European triangulation systems, *Bull. Geod.*, **56,** 177–190, 1960.

———, *Geodesy*, 2d edition, Oxford University Press, London, 561 pp., 1962.

BROUWER, D., Solution of the problem of artificial satellite theory without drag, *Astron. J.*, **64,** 378–397, 1959.

———, and CLEMENCE, G. M., *Methods of Celestial Mechanics*, Academic Press, New York, 598 pp., 1961.

BROWN, D. C., A matrix treatment of the general problem of least squares considering correlated observations, *Ballistic Res. Labs. Rept.*, **937,** 28 pp., 1955.

———, A treatment of analytical photogrammetry with emphasis on ballistic camera applications, *RCA Data Reduction Tech. Rept.*, **39,** Patrick AFB, Fla., 147 pp., 1957.

BROWN, W. F., Jr., Minimum variance in gravity analysis, *Geophysics*, **20,** 807–828, 1955; **21,** 107–141, 1956.

BULLARD, E. C., and COOPER, R. I. B., The determination of the masses necessary to produce a given gravitational field, *Proc. Roy. Soc. London, A*, **194,** 332–347, 1948.

CAPUTO, M., Gravity in space and the dimensions and mass of the earth, *J. Geophys. Res.*, **68,** 4595–4600, 1963.

CHAPMAN, S., and BARTELS, J., *Geomagnetism, vol.* 2, Oxford University Press, London, 1049 pp., 1940.

COOK, A. H., The calculation of deflection of the vertical from gravity anomalies, *Proc. Roy. Soc. London, A*, **204,** 374–395, 1950.

———, A note on the errors involved in the calculation of elevations of the geoid, *Proc. Roy. Soc. London, A*, **208,** 133–141, 1951.

———, The external gravity field of a rotating spheroid to the order of e^3, *Geophys. J.*, **2,** 199–214, 1959.

———, Report on absolute measurements of gravity, *Bur. Gravimetrique Intern., Assoc. Intern. Geod., Bull. Inform.*, **4,** 57–67, 1963a.

———, Geodetic constants and the motion of the moon, *Proc. IAU Symp.*, **21,** *Astron. Constants, Bull. Astron.*, in press, 1963b.

CORON, S., Anomalies isostatiques (Airy-30 km), Europe-Africa, *Bur. Gravimetrique Intern.*, Paris, 1962.

DEAN, W. C., Frequency analysis for gravity and magnetic interpretation, *Geophysics*, **23,** 97–127, 1958.

DE GRAAF-HUNTER, J., The figure of the earth from gravity observations and the precision obtainable, *Phil. Trans. Roy. Soc. London, A*, **234,** 377–431, 1935.

———, The shape of the earth's surface expressed in terms of gravity at ground level, *Bull. Geod.*, **56,** 191–200, 1960.

DESITTER, W., On the system of astronomical constants, *Bull. Astron. Inst. Neth.*, **8,** 213–231, 1938.

DURBIN, W. P., Jr., Some correlations of gravity and geology, *Proc. Symp. Geodesy Space Age, Publ. Inst. Geod. Phot. Cart.*, **15,** 130–135, 1961.

FISCHER, I., The Hough ellipsoid, *Bull. Geol.*, **54,** 45–52, 1959a.

———, A tentative world datum from geoidal heights based on the Haugh ellipsoid and the Columbus geoid, *J. Geophys. Res.*, **64,** 73–84, 1959b.

———, A map of geoidal contours in North America, *Bull. Geod.*, **57,** 321–324, 1960a.

———, The present extent of the astrogeodetic geoid and the geodetic world datum derived from it, *J. Geophys. Res.*, **65,** 2067–2076, 1960b.

———, The present extent of the astrogeodetic geoid and the geodetic world datum derived from it, *Bull. Geod.*, **61,** 245–264, 1961.

———, The parallax of the moon in terms of a world geodetic system, *Astron. J.*, **67,** 373–378, 1962.

FOUGERE, P. F., Spherical harmonic analysis, 1, A new method and its verification, *J. Geophys. Res.*, **68,** 1131–1139, 1963.

GARLAND, G. D. Gravity and isostasy, *Handbuch der Physik*, **47,** pp. 202–245, Springer-Verlag, Berlin, 1956.

GRANT, F., A problem in the analysis of geophysical data, *Geophysics*, **22,** 309–344, 1957.

GRANT, F. S., and ELSAHARTY, A. F., Bouguer gravity corrections using a variable density, *Geophysics*, **27,** 616–626, 1962.

GRENANDER, U., and ROSENBLATT, M., *Statistical Analysis of Stationary Time Series*, John Wiley & Sons, New York, 300 pp., 1957.

GROVES, G. V., Motion of a satellite in the earth's gravitational field, *Proc. Roy. Soc. London, A*, **254,** 48–65, 1960.

GUIER, W. H., Determination of the non-zonal harmonics of the geopotential from satellite Doppler data, *Nature*, in press, 1963.

HAMILTON, T. W., CAIN, D. L., SJOGREN, W. L., and NULL, G., Earth-moon system constants, Jet Propulsion Lab. preprint, 1963.

HARRISON, J. C., The measurement of gravity, *Proc. IRE*, **50,** 2302–2312, 1962.

HEISKANEN, W. A., The Columbus geoid, *Trans. Am. Geophys. Union*, **38,** 841–848, 1957.

———, On the activity of the isostatic institute of the IAG and of the worldwide gravity project of Columbus, Ohio, in physical geodesy, *Bull. Geod.*, **63,** 95–104, 1962.

———, and VENING-MEINESZ, F. A., *The Earth and its Gravity Field*, McGraw-Hill Book Co., New York, 470 pp., 1958.

HELMERT, F. R., *Die Mathematischen and Physikalischen Theorieen der höheren Geodasie*, **2**, *Die Physikalischen Theorie*, B. G. Teubner, Leipzig, 610 pp., 1884.

HIRVONEN, R. A., The continental undulations of the geoid, *Veroeff. Finn. Geod. Inst.*, **19**, 89 pp., 1934.

———, Gravity anomalies and deflections of the vertical above sea level, *Trans. Am. Geophys. Union*, **33**, 801–809, 1952.

———, On the precision of the gravimetric determination of the geoid, *Trans. Am. Geophys. Union*, **37**, 1–8, 1956.

———, New theory of gravimetric geodesy, *Publ. Isos. Inst. Intern. Assoc. Geod.*, **32**, Helsinki, 50 pp., 1960.

———, The reformation of geodesy, *J. Geophys. Res.*, **66**, 1471–1478, 1961 (*Bull. Geod.*, **65**, 197–214, 1962).

———, On the statistical analysis of gravity anomalies, *Pub. Isos. Inst. Intern. Assoc. Geod.*, **37**, Helsinki, 24 pp., 1962.

HOLLOWAY, J. L., Jr., Smoothing and filtering of time series and space fields, *Advan. Geophys.*, **4**, 351–389, 1958.

IZSAK, I. G., A theory of satellite motion about an oblate planet, 1, A second-order solution of Vinti's dynamical problem, *Smithsonian Inst. Astrophys. Obs. Spec. Rept.*, **52**, 54 pp., 1960.

———, A determination of the ellipticity of the earth's equator from the motion of two satellites, *Astron. J.*, **66**, 226–229, 1961a.

———, Differential orbit improvement with the use of rotated residuals, *Smithsonian Inst. Astrophys. Obs. Spec. Rept.*, **73**, 9 pp., 1961b.

———, On the critical inclination in satellite theory, *Smithsonian Inst. Astrophys. Obs. Spec. Rept.*, **90**, 24 pp., 1962.

———, Tesseral harmonics in the geopotential, *Nature*, **199**, 137–139, 1963.

JEFFREYS, H., An application of the free-air reduction of gravity, *Gerlands Beitr. Geophys.*, **31**, 378–386, 1931.

———, The determination of the earth's gravitational field (second paper), *Monthly Notices Roy. Astron. Soc.*, **5**, 55–66, 1943.

———, *The Earth*, 4th edition, Cambridge University Press, London, 420 pp., 1959.

———, The analysis of gravity, *Smithsonian Inst. Astrophys. Obs. Spec. Rept.*, **79**, 12 pp., 1961.

JONES, R. H., Spectral analysis with regularly missed observations, *Ann. Math. Stat.*, **33**, 455–461, 1962.

———, Stochastic processes on a sphere, *Ann. Math. Stat.*, **34**, 213–218, 1963.

JUNG, K., Die rechnerische Behandlung der Airyschen Isostasie mit einer Entwichlung des Quadrats der Meereshohen nach Kugel funktionen, *Gerlands Beitr. Geophys.*, **62**, 39–56, 1952.

———, Figur der Erde, *Handbuch der Physik, Geophysik*, **1**, pp. 534–639, Springer-Verlag, Berlin, 1956.

KALMAN, R. E., A new approach to linear filtering and prediction problems, *Trans. ASME, D, J. Basic Eng.*, **82**, 35–45, 1960.

KÄRKI, P., KIVIOJA, L., and HEISKANEN, W. A., Topographic-isostatic reduction maps for the world for the Hayford zones 18–1, Airy-Heiskanen system, $T=30$ km, *Publ. Isos. Inst. Intern. Assoc. Geod.*, **35**, 25 pp., 1961.

KAULA, W. M., Accuracy of gravimetrically computed deflections of the vertical, *Trans. Am. Geophys. Union*, **39**, 1027–1033, 1957.

———, Statistical and harmonic analysis of gravity, *Army Map Serv. Tech. Rept.*, **24**, 141 pp., 1959a.

———, Statistical and harmonic analysis of gravity, *J. Geophys. Res.*, **64**, 2401–2421, 1959b.

———, Analysis of gravitational and geometric aspects of geodetic utilization of satellites, *Geophys. J.*, **5**, 104–133, 1961a.

———, A geoid and world geodetic system based on a combination of gravimetric, astrogeodetic, and satellite data, *J. Geophys. Res.*, **66**, 1799–1812, 1961b.

———, Analysis of satellite observations for longitudinal variations of the gravitational field, in *Space Research, Proc. Intern. Space Sci. Symp., 2nd, Florence*, pp. 360–372, North-Holland Publishing Co., Amsterdam, 1961c.

———, Celestial geodesy, *Advan. Geophys.*, **9**, 191–293, 1962.

———, Satellite orbit analysis for geodetic purposes, in *Proc. IUTAM Symp. Dynamics of Satellites, Paris*, pp. 205–210, Springer-Verlag, Berlin, 1963a.

———, Tesseral harmonics of the gravitational field and geodetic datum shifts derived from camera observations of satellites, *J. Geophys. Res.*, **68**, 473–484, 1963b.

———, A review of geodetic parameters, in *Proc. IAU Symp.*, **21**, *Astron. Constants, Bull. Astron.*, in press, 1963c.

———, Improved geodetic results from camera observations of satellites, *J. Geophys. Res.*, **68**, 5183–5190, 1963d.

———, and FISCHER, I., U.S. Army world geodetic system 1959, 1, Methods, *Army Map Serv. Tech. Rept.*, **27**, 62 pp., Washington, 1959.

KING-HELE, D. G., The effect of the earth's oblateness on the orbit of a near satellite, *Proc. Roy. Soc. London, A*, **241**, 49–72, 1958.

———, COOK, G. E., and REES, J. M., Determination of the even harmonics in the earth's gravitational potential, *Geophys. J.*, in press, 1963.

KIVIOJA, L., Development of gravity Bouguer anomalies of state of Ohio and the isostatic anomalies in North Atlantic in Fourier series, *Inst. Geod. Phot. Cart. Rept.*, **22**, 13 pp., 1962.

KOZAI, Y., On the effects of the sun and the moon upon the motion of a close earth satellite, *Smithsonian Inst. Astrophys. Obs. Spec. Rept.*, **22**, 7–10, 1959a.

———, The motion of a close earth satellite, *Astron. J.*, **64**, 367–377, 1959b.

———, Tesseral harmonics of the potential of the earth as derived from satellite motions, *Astron. J.*, **66**, 355–358, 1961.

———, Numerical results from orbits, *Smithsonian Inst. Astrophys. Obs. Spec. Rept.*, **101**, 19 pp., 1962.

———, The potential of the earth derived from satellite motions, in *Proc. IUTAM Symp. Dynamics of Satellites, Paris*, pp. 65–73, Springer-Verlag, Berlin, 1963a.

———, Review paper (in Japanese), *Japanese Astron. Soc. Astron. Herald*, 1963b.

KRUMBEIN, W. C., Trend surface analysis of contour-type maps with irregular control-point spacing, *J. Geophys. Res.*, **64**, 823–834, 1959.

LACOSTE, L. B., Jr., and HARRISON, J. C., Gravimetry, *Rev. Geophys.*, in preparation, 1964.

LAMBERT, W. D., The gravity field of an ellipsoid of revolution as a level surface, *Publ. Inst. Geod. Phot. Cart.*, **14**, 42 pp., 1961.

LEDERSTEGER, K., Der physikalische Zusammenhang Zwischen der statischen Abplattung und dem inneren Aufbau der Erde der hypothetischen Grundlage einer sechsparametrigen Gleichgewichtsfigur, *Geofis. Pura Appl.*, **51**, 1–28, 1962.

LEVALLOIS, J. J., Sur la formule de Stokes et celles qui en dérivent, *Festschrift C. F. Baeschlin*, pp. 165–183, Orell Füssli Verlag, Zurich, 1957.

———, Considerations generales sur les réductions de la pesanteur, *Bull. Geod.*, **63**, 79–93, 1962.

MCCAHAN, A. L., Summary of progress in gravity activities in the United States, Presentation to the Conference Com. Survey Off., Cambridge, 1963.

MACOMBER, M., Project ANNA, in *COSPAR-IAG Symp. Geod. Use of Satellites*, North-Holland Publishing Co., Amsterdam, 257–260, 1963.

MANDELBAUM, H., Statistical and geological implications of trend mapping with nonorthogonal polynomials, *J. Geophys. Res.*, **68**, 505–520, 1963.

MERSON, R. H., The motion of a satellite in an axisymmetric gravitational field, *Geophys. J.*, **4**, 17–52, 1961.

———, A Pegasus computer programme for the improvement of the orbital parameters of an earth satellite, in *Proc. IUTAM Symp. Dynamics of Satellites, Paris*, pp. 83–110, Springer-Verlag, Berlin, 1963.

MOE, K., On the optimum choice of a smoothing interval for computing definitive orbits of artificial satellites, in *COSPAR-IAG Symp. Geod. Use Satellites*, North-Holland Publishing Co., Amsterdam, 232–241, 1963.

MOLODENSKIY, M. S., Basic problems of geodetic gravimetry (in Russian), *Tr. TsNIIGAik*, **42**, Moscow, 107 pp., 1945 (partially republished as *Dept. Com. Office Tech. Serv. Trans.* **58–11**, 257, 1959).

———, The external gravitational field and the figure of the physical surface of the earth (in Russian), *Izv. Akad. Nauk SSSR, Ser. Geogr. Geofiz*, **12**, 193–211, 1948.

———, Lösung der Integral Gleichung zur Ermittlung der Erdfigur, *Bull. Geod.*, **64**, 181–187, 1962.

———, YEREMEYEV, V. F., and YURKINA, M. I., Methods for the study of the external gravitational field and figure of the earth (in Russian), *Tr. Central Inst. Geod. Phot. Cart., Moscow*, **131**, 251 pp., 1960 (translation published by Dept. Com. Office Tech. Serv., Washington, 1962).

———, ———, and ———, An evaluation of accuracy of Stokes' series and of some attempts to improve his theory, *Bull. Geod.*, **63**, 19–38, 1962.

MORANDO, B., Orbites de resonance des satellites de 24 h., *Bull. Astron.*, **24**, 47–67, 1963.

MORELLI, C., Report of the special study group 5 on the absolute and first order world net, *Bur. Gravimetrique Intern., Assoc. Intern. Geod., Bull. Inform.*, **4**, 70–77, 1963.

MORITZ, H., Eine Integral Gleichung des Geoides, *Gerlands Beitr. Geophys.*, **70**, 373–379, 1961.

———, Studies on the accuracy of the computation of gravity in high elevations, *Publ. Isos. Inst. Intern. Assoc. Geod.*, **38**, Helsinki, 37 pp., 1962a.

———, Interpolation and prediction of gravity and their accuracy, *Inst. Geod. Phot. Cart. Rept.*, **24**, 69 pp., 1962b.

———, A statistical method for upward and downward continuation of gravity; on the arrangement of gravity stations for the upward continuation, *Inst. Geod., Phot., Cart. Rept.*, **25**, 27 pp., 1962c.

MUELLER, I. I., *Introduction to Satellite Geodesy*, F. Ungar Publishing Co., New York, in press, 1963.

MUNK, W. H., and MACDONALD, G. J. F., *The Rotation of the Earth*, Cambridge University Press, New York, 323 pp., 1960.

MUSEN, P., Application of Hansen's theory to the motion of an artificial satellite in the gravitational field of the earth, *J. Geophys. Res.*, **64**, 2271–2279, 1959.

———, On the motion of a satellite in an asymmetrical gravitational field, *J. Geophys. Res.*, **65**, 2783–2792, 1960.

———, and BAILIE, A., On the motion of a 24-hour satellite, *J. Geophys. Res.*, **67**, 1123–1132, 1962.

NEWTON, R. R., Ellipticity of the equator deduced from the motion of Transit 4A, *J. Geophys. Res.*, **67**, 415–416, 1962.

———, HOPFIELD, H. S., and KLINE, R. C., Odd harmonics of the earth's gravitational field, *Nature*, **190**, 617–618, 1961.

OBUKHOV, A. M., Statistically homogeneous random fields on a sphere (in Russian), *Usp. Mat. Nauk*, **2**, 196–198, 1947.

O'KEEFE, J. A., ECKELS, A., and SQUIRES, R. K., The gravitational field of the earth, *Astron. J.*, **64**, 245–253, 1959.

OLDHAM, C. H. G., and SUTHERLAND, D. B., Orthogonal polynomials: their use in estimating the regional effect, *Geophysics*, **20**, 295–306, 1955.

OLIVER, H. W., PAKISER, L. C., and KANE, M. F., Gravity anomalies in the central Sierra Nevada, Calif., *J. Geophys. Res.*, **66**, 4265–4271, 1961.

ORLIN, H., The three components of the external anomalous gravity field, *J. Geophys. Res.*, **64,** 2393–2399, 1959.

PARZEN, E., An approach to time series analysis, *Ann. Math. Stat.*, **32,** 951–989, 1961.

———, Extraction and detection problems, probability density functionals, and reproducing kernel Hilbert spaces, *Appl. Math. Stat. Lab. Tech. Rept.*, Stanford, 57 pp., 1962a.

———, On spectral analysis with missing observations and amplitude modification, *Appl. Math. Stat. Lab. Tech. Rept.*, **46,** Stanford, 19 pp., 1962b.

PELLINEN, L. P., Accounting for topography in the calculation of quasi-geoidal heights and plumb line deflections from gravity anomalies, *Bull. Geod.*, **63,** 57–65, 1962.

PETERS, L. J., The direct approach to magnetic interpretation and its practical application, *Geophysics*, **14,** 290–320, 1949.

PIZZETTI, P., Sulla espressione della gravitá alla superficie del geoide, supposto ellissoidico, *Atti Accad. Naz. Lincei*, **3,** 166, 1894.

PRESS, F., Crustal structure in the California-Nevada region, *J. Geophys. Res.*, **65,** 1039–1051, 1960.

RAPP, R., Correlation coefficients and their use in the prediction of mean anomalies, *Inst. Geod. Phot. Cart. Rept.*, **20,** 39 pp., 1962.

RICE, D. A., Deflections of the vertical from gravity anomalies, *Bull. Geod.*, **25,** 285–312, 1952.

———, A geoidal section in the United States, *Bull. Geod.*, **65,** 243–252, 1962.

ROBINSON, E. A., *An Introduction to Infinitely Many Variates*, C. Griffin & Co., London, 132 pp., 1959.

SAGREBIN, D. W., The theory of the regularized geoid (in Russian), *Publ. Inst. Theoret. Astron., Leningrad*, 1952 (German translation, *Geod. Inst. Potsdam*, **9,** 129 pp., 1956).

SCHOENBERG, I. J., Positive definite functions on spheres, *Duke Math. J.*, **9,** 96–108, 1942.

SHAPIRO, H. S., and SILVERMAN, R. A., Alias-free sampling of random noise, *J. Soc. Ind. Appl. Math.*, **8,** 225–248, 1960.

SHELKEY, M. J., The gravitational field of the earth; zonal harmonics from transit 1B and transit 2A data, *U.S. Naval Weapons Lab. Rept.*, **1807,** 13 pp., 1962.

SMITH, D. E., Determination of the earth's gravitational potential from satellite orbits, *Planet Space Sci.*, **8,** 45–48, 1961.

———, A determination of the odd harmonics in the geopotential function, *Planet. Space Sci.*, **11,** 789–796, 1963.

STEARN, J. L., and RICHARDSON, H., Adjustment of conditions with parameters and error analysis, *Bull. Geod.*, **64,** 117–136, 1962.

STOKES, G. G., On the variation of gravity at the surface of the earth, *Trans. Cambridge Phil. Soc. Math. Phys.*, **2,** 131–171, 1849.

STRAKHOV, V. N., The analytical continuation of two-dimensional potential fields with applications to the solution of the inverse problem of magnetic and gravitational exploration, *Izv. Acad. Nauk SSSR, Ser. Geofiz.*, **3,** 307–316 and 336–347; **4,** 491–505, 1962 (translation, *Bull. Akad. Sci. USSR, Geophys. Ser.*, **1962,** 209–214, 227–232, 224–331, 1962).

SZABO, B., Application of the gravimetric method for a world geodetic system, *Bull. Geod.*, **65,** 221–226, 1962.

TALWANI, M., WORZEL, J. L., and EWING, M., Gravity anomalies and crustal section across the Tonga trench, *J. Geophys. Res.*, **66,** 1265–1278, 1961.

TANNI, L., On the continental undulations of the geoid as determined from the present gravity material, *Publ. Isos. Inst. Intern. Assoc. Geod.*, **18,** 78 pp., 1948.

TENGSTRÖM, E., Calculation of the external gravity anomalies and deflections of the vertical at higher elevations by means of Taylor expansions from the geoid, *Inst. Geod. Phot. Cart. Rept.*, **5,** 14 pp., 1959.

———, A comparison between the methods of Stokes, Molodenskiy, and Hirvonen in physical geodesy, *Proc. Symp. Geod. Space Age, Publ. Inst. Geod. Phot. Cart.*, **15,** 148–154, 1961.

———, A simplified Rudzki reduction as a convenient method of reducing gravity data for geodetical and geophysical purposes, *Bull. Geod.*, **63,** 43–47, 1962.

TOMODA, Y., and ATI, K., Use of the function $\sin x/x$ in gravity problems, *Proc. Japan Acad.*, **31,** 443–448, 1955.

TSUBOI, C., Application of double Fourier series to computing gravity anomalies and other gravimetric quantities at any elevation from surface gravity anomalies, *Inst. Geod. Phot. Cart. Rept.*, **1,** 19 pp., and **2,** 16 pp., 1959.

UOTILA, U. A., Investigations on the gravity field and shape of the earth, *Publ. Inst. Geod. Phot. Cart.*, **10,** Columbus, 92 pp., 1960.

———, Existing gravity material, *Proc. Symp. Geod. Space Age, Publ. Inst. Geod. Phot. Cart.*, **15,** Columbus, 91–97, 1961.

———, Harmonic analysis of worldwide gravity material, *Publ. Isos. Inst. Intern. Assoc. Geod.*, **39,** Helsinki, 18 pp., 1962.

VAN ISACKER, J., Generalized harmonic analysis, *Advan. Geophys.*, **7,** 189–214, 1961.

VEIS, G., Geodetic uses of artificial satellites, *Smithsonian Contrib. Astrophys.*, **3,** 95–161, 1960.

———, and MOORE, C. H., Smithsonian Astrophysical Observatory Differential Orbit Improvement Program, *Jet Prop. Lab. Seminar Proc.: Tracking Programs and Orbit Determination*, 165–184, 1960.

VENING-MEINESZ, F. A., A formula expressing the deflection of the plumb-line in the gravity anomalies and some formulae for the gravity field and the gravity potential outside the geoid, *Verhandel-Koninkl. Ned. Akad. Wetenschap.*, **31,** 315–331, 1928.

———, The results of the development of the earth's topography in spherical harmonics up to the 31st order: provisional conclusions, *Koninkl. Ned. Akad. Wetenschap. Proc., B,* **62,** 115–125, 1959.

VINTI, J. P., New method of solution for unretarded satellite orbits, *J. Res. NBS, B,* **62,** 105–116, 1959.

———, Theory of an accurate intermediary orbit for satellite astronomy, *J. Res. NBS, B,* **65,** 169–202, 1961.

WOOLLARD, G. P., Crustal structure from gravity and seismic measurements, *J. Geophys. Res.*, **64,** 1521–1544, 1959.

———, The status of gravimetric control for global geodetic studies, *Proc. Symp. Geod. Space Age, Publ. Inst. Geod. Phot. Cart.,* **15,** Columbus, 97–115, 1961.

———, The relation of gravity anomalies to surface elevation, crustal structure, and geology, *Univ. Wisconsin Geophys. Polar Res. Center Res. Rept.,* **62–9,** 334 pp., 1962.

———, OSTENSO, N. A., THIEL, E., and BONINI, W. F., Gravity anomalies, crustal structure, and geology in Alaska, *J. Geophys. Res.,* **65,** 1021–1038, 1960.

———, and ROSE, J. C., *International Gravity Measurements,* Society of Exploration Geophysicists, Tulsa, 519 pp., 1963.

WORZEL, J. L., COLLETTE, B. J., SCHULZE, R., OKUDA, T., RICE, D. A., SLICHTER, L. B., and KARNICK, H., Report on gravity measurements at sea, *Bull. Inform. Bur. Grav. Intern.,* **4,** 91–98, 1963.

YAGLOM, A. M., Second order homogeneous random fields, *Proc. 4th Berkeley Sym. Math. Stat. Prov., vol.* **2,** pp. 593–622, University of California Press, Berkeley, 1961.

YAPLEE, B. S., KNOWLES, S. H., SHAPIRO, A., CRAIG, K. J., and BROUWER, D., The mean distance to the moon as determined by radar, *Proc. IAU Symp.,* **21,** *Astron. Constants, Bull. Astron.,* in press, 1963.

ZHONGLOVITCH, I. D., The external gravity field of the earth and the fundamental constants connected with it (in Russian), *Acad. Sci. Publ. Inst. Teor. Astron.,* Leningrad, 129 pp., 1952.

———, Determination of the dimensions of the general earth ellipsoid (in Russian), *Trud. Inst. Teor. Astron.,* **6,** Leningrad, 5–66, 1956.

IMPROVED GEODETIC RESULTS FROM CAMERA OBSERVATIONS OF SATELLITES

W. M. KAULA

Goddard Space Flight Center

Analysis of Baker Nunn camera observations of satellites $1959\alpha_1$ over 1032 days; 1959η over 792 days; $1960\iota_2$ over 480 days; $1961\delta_1$ over 150 days; and $1961\alpha\delta_1$ over 54 days yielded results greatly improved over those previously reported. This improvement is primarily due to the use of much more data and secondarily to various modifications in the method of analysis. As indicated by the discrepancies between results from appreciably different orbits, the datum shifts obtained have standard deviations of ± 4 to ± 23 meters. The tesseral harmonic of the gravity field most firmly determined appears to be J_{41}, followed by J_{22}, J_{31}, J_{42}, J_{43}, and J_{32}. The principal sources of error suggested are the influence of preassigned variances on separation of gravitational coefficients having the same periodic effects on an orbit (e.g., J_{22} and J_{42}) and the holding fixed with respect to each other stations on the same geodetic datum, but they do not seem adequate to explain all systematic discrepancies in the results. A comparison of gravitational and geometric geoid heights at station positions gave a mean equatorial radius of $6{,}378{,}196 \pm 11$ m.

INTRODUCTION

This paper describes appreciable improvement over the results obtained by Kaula (1963a), due both to additional observations and to revisions in the method of analysis. The discussion in this paper follows closely that of Kaula (1963a) and is limited to changes therefrom.

OBSERVATIONS

The precisely reduced Baker Nunn camera observations of $1959\alpha_1$, 1959η, and $1960\iota_2$ from launch until the end of 1961, of $1961\delta_1$ from launch until the middle of 1961, and of $1961\alpha\delta_1$ in the spring of 1962 were analyzed. The observations through mid-1961 have been published in the catalogs compiled by Veis et al. (1961–62).

No change was made in the methods of selection and conversion of observations. The number of observations of each satellite used is given in table 1.

GEOMETRY

The initial station positions used were the solutions given in table 1 of Kaula (1963a) with corrections for errors in the computed positions of three stations relative to the principal datums provided by I. G. Izsak of the Smithsonian Institution Astrophysical Observatory. Corrections to coordinates u_1, u_2, and u_3 in earth radii are listed in sequence (all values times 10^{-6}):

San Fernando	+5.6	−5.0	−8.5
Naini Tal	+2.7	−5.0	+4.9
Curaçao	−1.0	0.0	+2.8

The datum shifts listed in table 2 of the present paper apply to the starting coordinates in column 4 of table 1 of Kaula (1963a) with the local corrections given above.

DYNAMICS

The only change made in the dynamical aspects of the treatment was to omit entirely the long-period and secular perturbations that are due to lunisolar attraction, radiation pressure, and drag by a specified atmospheric model. For the orbital arc lengths of 10 to 20 days it was found that these effects were adequately absorbed by an arbitrary acceleration in the mean anomaly. Their inclusion made little difference in the solutions

TABLE 1.—*Satellite Orbit Specifications*

	Satellite				
	$1959\alpha_1$	1959η	$1960\iota_2$	$1961\delta_1$	$1961\alpha\delta_1$
Epoch	1959 Feb. 28.5	1959 Sept. 28.5	1960 Sept. 22.0	1961 Feb. 20.0	1962 Mar. 8.5
Semimajor axis	1.304585	1.334500	1.250057	1.252779	1.568136
Eccentricity	0.16582	0.19008	0.01146	0.12135	0.01197
Inclination	0.57381	0.58212	0.82434	0.67835	1.67316
Argument of perigee	3.36062	3.20403	2.26377	2.02733	4.28853
Longitude of node	2.52442	3.48304	2.28139	2.76786	5.71336
Mean anomaly	6.00463	3.82408	2.72868	5.96587	1.51124
Perigee motion/day	+0.09181	+0.08501	+0.05186	+0.08315	−0.01733
Node motion/day	−0.06108	−0.05712	−0.05413	−0.06347	+0.00367
Max. A/m, cm²/g	0.21	0.27	0.27	15.9	0.08
Min. A/m, cm²/g	0.21	0.04	0.08	15.9	0.02
Perigee height, km	560	510	1500	640	3500
Number of days	1032	792	480	150	54
Number of observations	3513	3034	2502	1395	552

TABLE 2.—*Datum Shifts (in length units of* 6.378165 *m)*

Datum	Coordinate	$1959\alpha_1$	1959η	$1960\iota_2$	$1961\delta_1$	$1961\alpha\delta_1$	Weighted mean
Americas	Δu_1	−02.5	−02.6	−03.7	−03.8	−06.4	−03.8±1.0
	Δu_2	−04.7	−05.2	−09.6	−11.3	−04.2	−05.1±0.8
	Δu_3	−00.9	−00.5	−00.6	−01.9	−00.3	−00.4±0.2
Europe–Africa	Δu_1	+06.5	+07.3	+06.6	+11.6	+04.6	+05.8±0.7
Siberia–India	Δu_2	−07.8	−07.8	−09.3	−04.1	−10.2	−08.9±0.5
	Δu_3	+02.0	+01.3	+02.2	−01.4	+02.1	+01.9±0.2
Australia	Δu_1	−16.3	−19.6	−19.6	−11.2	−26.6	−17.3±1.5
	Δu_2	+09.6	+06.0	+06.7	+07.5	+03.0	+05.2±1.7
	Δu_3	+10.8	+14.6	+10.0	+14.7	+09.4	+10.5±0.4
Japan–Korea–Manchuria	Δu_1	−08.9	−11.5	−08.3	−08.5	−06.5	−08.9±0.5
	Δu_2	+04.1	+05.2	+13.0	+08.7	+09.3	+09.4±0.7
	Δu_3	+01.4	+00.1	+01.8	−00.1	+04.4	+01.5±0.8
Argentina	Δu_1	+35.6	+37.9	+39.9	+50.7	+34.4	+38.3±1.6
	Δu_2	−03.7	+03.8	−02.4	−02.1	+00.0	−02.3±0.6
	Δu_3	+10.0	+07.5	+09.9	+04.2	−06.3	+05.7±3.5
Hawaii	Δu_1	+03.3	+01.3	−05.2	+00.4	−00.3	−04.0±1.6
	Δu_2	+06.1	+04.5	+16.0	+01.2	+15.2	+09.2±2.8
	Δu_3	−45.4	−48.6	−47.7	−67.9	−25.0	−45.5±3.6

obtained for tesseral harmonics or station shifts—if anything, they may have distorted the results by shifting computed satellite directions farther from those observed.

DATA ANALYSIS

Of the five techniques listed on page 478 of Kaula (1963a) only three were employed: preassigning a covariance matrix **V** for the starting values of parameters, assigning higher weight to the across-track than to the along-track component of an observation, and using arbitrary polynomials.

The covariances and variances preassigned were identical with those in Kaula (1963a) with the exception of $\bar{C}_{21}$, $\bar{S}_{21}$, which were held fixed.

The observational variance employed was $(0.026\ \text{sec})^2$ time and $(9.2\ \text{sec})^2$ direction for 12-day arcs of $1959\alpha_1$ and 1959η; $(0.047\ \text{sec})^2$ time and $(13.4\ \text{sec})^2$ direction for 20-day arcs of $1960\iota_2$; $(0.146\ \text{sec})^2$ time and $(43.8\ \text{sec})^2$ direction for 10-

TABLE 3.—*Gravitational Coefficient Solutions. Multiply All Numbers by a Scaling Factor of* 10^{-6}

Coefficient*	$1959\alpha_1$	1959η	$1960\iota_2$	$1961\delta_1$	$1961\alpha\delta_1$	Weighted mean
$\Delta\bar{C}_{00}$	4.96	−8.88	−0.75	−18.50	−9.85	−2.46±2.36
$\Delta\bar{C}_{20}$	−0.06	−0.06	−0.05	−0.29	0.00	−0.03±0.02
$\bar{C}_{22}$	1.30	1.36	1.99	1.80	2.52	1.88±0.29
$\bar{S}_{22}$	−1.74	−0.76	−1.63	−0.32	−0.89	−1.38±0.17
$\bar{C}_{30}$	0.97	0.96	0.98	1.01	0.97	0.97±0.01
$\bar{C}_{31}$	1.30	1.62	1.53	−0.96	1.18	1.52±0.03
$\bar{S}_{31}$	0.29	0.99	−0.10	−0.34	0.46	0.14±0.16
$\bar{C}_{32}$	−0.14	−0.13	0.29	2.35	−0.84	−0.02±0.26
$\bar{S}_{32}$	0.49	0.29	0.38	−0.16	0.98	0.42±0.06
$\bar{C}_{33}$	0.36	1.11	0.42	2.36	1.70	0.70±0.26
$\bar{S}_{33}$	0.83	1.11	0.89	0.43	−1.33	0.76±0.29
$\bar{C}_{40}$	0.68	0.67	0.61	−0.35	0.62	0.67±0.02
$\bar{C}_{41}$	−0.38	−0.38	−0.33	−0.48	−1.00	−0.33±0.01
$\bar{S}_{41}$	0.43	0.53	0.45	0.43	−0.45	0.37±0.15
$\bar{C}_{42}$	−0.10	−0.10	0.02	0.03	0.47	0.01±0.02
$\bar{S}_{42}$	0.52	0.68	0.36	−0.43	0.06	0.35±0.15
$\bar{C}_{43}$	0.18	0.35	0.50	0.44	0.17	0.17±0.02
$\bar{S}_{43}$	0.29	0.11	−0.00	0.16	0.42	0.41±0.03
$\bar{C}_{44}$	0.12	0.01	−0.20	0.20	−0.24	−0.01±0.08
$\bar{S}_{44}$	0.11	0.22	0.36	0.29	0.32	0.18±0.05
$\bar{C}_{50}$	0.02	0.03	0.03	0.01	0.02	0.02±0.01
$\bar{C}_{51}$	−0.14	−0.02	−0.01	−0.63	(†)	−0.13±0.02
$\bar{S}_{51}$	−0.06	−0.03	−0.01	0.23	(†)	−0.01±0.01
$\bar{C}_{60}$	−0.09	−0.08	−0.04	1.10	−0.10	−0.09±0.02
$\bar{C}_{61}$	−0.01	−0.03	0.00	−0.26	−0.09	−0.05±0.03
$\bar{S}_{61}$	−0.09	−0.02	−0.07	−0.49	−0.06	−0.06±0.01
$\bar{C}_{62}$	−0.04	0.05	−0.01	−0.07	0.05	0.01±0.01
$\bar{S}_{62}$	−0.09	−0.18	−0.01	−0.07	0.01	−0.02±0.03
$\bar{C}_{63}$	−0.02	−0.10	0.15	−0.02	(†)	0.15±0.01
$\bar{S}_{63}$	−0.12	−0.01	−0.08	−0.06	(†)	−0.08±0.01
$\bar{C}_{64}$	−0.00	0.06	0.06	−0.19	−0.01	−0.01±0.01
$\bar{S}_{64}$	−0.06	−0.09	−0.42	−0.30	0.03	−0.03±0.07
$\bar{C}_{70}$	0.12	0.12	0.07	0.09	0.12	0.12±0.01

*$\bar{C}_{nm}$ and $\bar{S}_{nm}$ are coefficients of spherical harmonic terms $kM/r(a/r)_nH_{nm}$ such that $\int H_{nm}^2\, d\sigma = 4\pi$ for integration over the sphere. $\Delta\bar{C}_{00}$ and $\Delta\bar{C}_{20}$ are corrections to 0.3986032×10^{21} $(1.0-0.00108236P_2)$ cgs.

†No determinations of $\bar{C}_{51}$, $\bar{S}_{51}$, $\bar{C}_{63}$, $\bar{S}_{63}$ were made from $1961\alpha\delta_1$ because the partial derivatives of the orbit with respect to these coefficients were all smaller than the criterion $0.1n^{1.2}$ (Kaula, 1963a).

day arcs of $1961\delta_1$; and $(0.047 \text{ sec})^2$ time and $(13.4 \text{ sec})^2$ direction for 18-day arcs of $1961\alpha\delta_1$. The principal criterion used in determining the observational variances was the χ^2 test; i.e., the quantity

$$s = (\mathbf{f}^T\mathbf{W}^{-1}\mathbf{f} - \mathbf{z}^T\mathbf{M}^T\mathbf{W}^{-1}\mathbf{f})/(n-p) \qquad (1)$$

should average 1 for several orbital arcs, where $\mathbf{f}$ is the vector of observation equation residuals; $\mathbf{W}$ is the covariance matrix of observations; $\mathbf{z}$ is the vector of corrections to parameters; $\mathbf{M}$ is the matrix parameter coefficients in the observation equations; n is the number of observations; and p is the number of free parameters. In forming the covariance matrix $\mathbf{W}$, observations in the same pass were treated as having the same timing error.

The arc lengths used were chosen after some experimentation as giving a reasonable compromise between magnitude of residuals and number of observations.

The use of arbitrary polynomials was held to a minimum; i.e., the only one used was a t^2 variation in the mean anomaly.

In determining the estimated mean value and its standard deviation from several orbital arcs of the same satellite, the weighting of a particular arc was considered to be proportionate to its degrees of freedom. The computer program limited to fifteen the number of arcs that could be combined at a time. In combining the results of several sets of fifteen (or fewer) arcs, the weight ascribed to the mean of each set was considered to be the inverse of its variance, or standard deviation squared.

In order that the final mean and standard deviation reflect as much as possible any systematic differences which were functions of orbital specifications, all sets were combined, with inverse-variance weighting, into four groups: $1959\alpha_1$ and 1959η, twelve sets; $1960\iota_2$, two sets; $1961\delta_1$, one set; and $1961\alpha\delta_1$, one set. The final means and standard deviations given in tables 2 and 3 are the result of an inverse-variance weighted combination of these four group solutions. However, for most of the variables, the standard deviations from combining the four groups were smaller than the standard deviations combining all sixteen sets at once, primarily because the differences between the $1960\iota_2$ mean and the $1959\alpha_1$ and 1959η mean were smaller than the scatter of $1959\alpha_1$ and 1959η solutions about their own mean.

To avoid the tendency to prejudge the order of magnitude of the solution, which is the main defect of the preassigned-variance technique, some computer experimentation was tried in determining the amplitudes of specified periodic variations, in place of harmonic coefficients, in holding the reference orbit fixed, and in analyzing residuals. Applying these methods to one satellite at a time did not give as good results as the preassigned-variance method, to judge by the scatter of solutions. To apply them to data from more than one satellite simultaneously required considerable program revision which did not seem worth while because this method has been applied extensively by Izsak (1963). Other changes tried and dropped as unnecessary were deleting orbital segments for which observations are scanty and holding fixed the station shifts obtained from the previous analysis of $1960\iota_2$ observations. Also dropped was the device of weighting observations inversely as their density with respect to the phase angle (node-GST).

RESULTS

The analysis described above took much time to apply to the large quantity of $1959\alpha_1$ and 1959η data. The attempt to combine solutions from different sets of arcs was not made until this analysis had been completed. Consequently, the good agreement shown by tables 2 and 3 between the results from $1960\iota_2$ on the one hand and from $1959\alpha_1$ and 1959η on the other came as a pleasant surprise. The combination of results is not as good, of course, as is suggested by the formal standard deviations given in the tables; in particular, the errors in difference of position between stations in North and South America—or between stations in Europe, Africa, and India—which were held fixed with respect to each other, are probably several times as great as some of the stated uncertainties. The good agreement is even more marked for the spatial representations given in figures 1 and 2; e.g., for the seven most extreme maximums and minimums in the Vanguard geoid of figure 1, there are maximums and

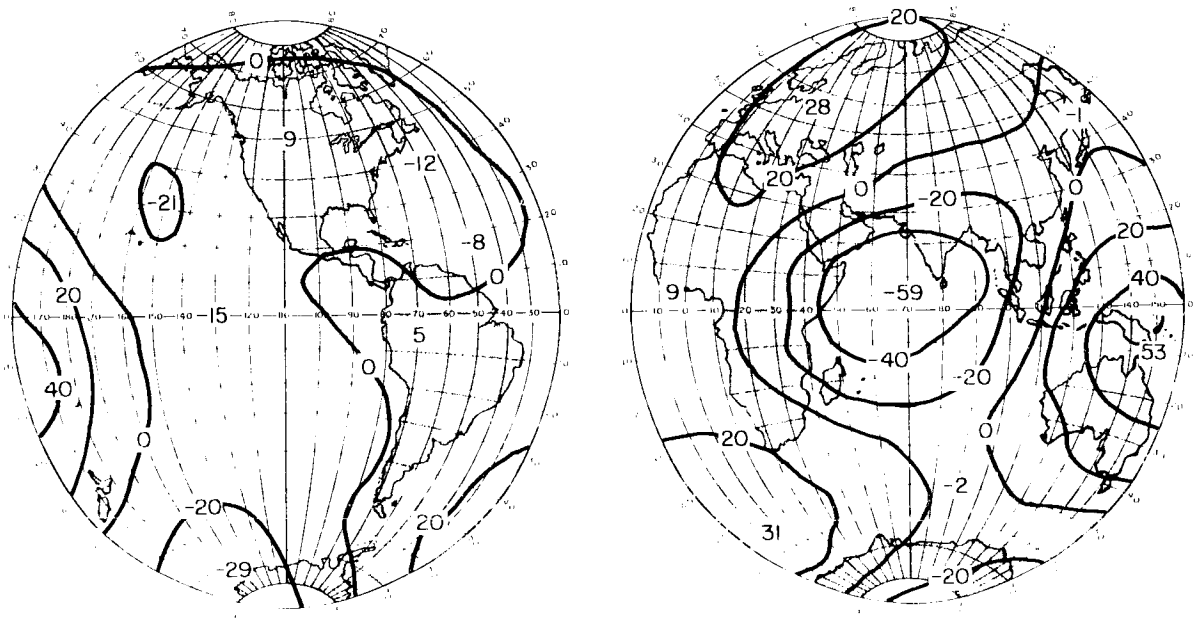


FIGURE 1.— Vanguard geoid. Geoid heights, in meters, referred to an ellipsoid of flattening 1/298.24, determined from observations of satellites, $1959\alpha_1$ and 1959η.

minimums in the Echo rocket geoid of figure 2 agreeing within 10° in location and within 11 m in magnitude. The degree of independence in these solutions is fairly satisfying. The orbits differ by 0.23 in inclination, and 0.16 in eccentricity, the arc lengths used differed in a ratio of 5 to 3, and the observational weighting differed in a ratio of 3 to 2. It would be very desirable, however, to obtain comparable series of observations of a satellite of much higher inclination.

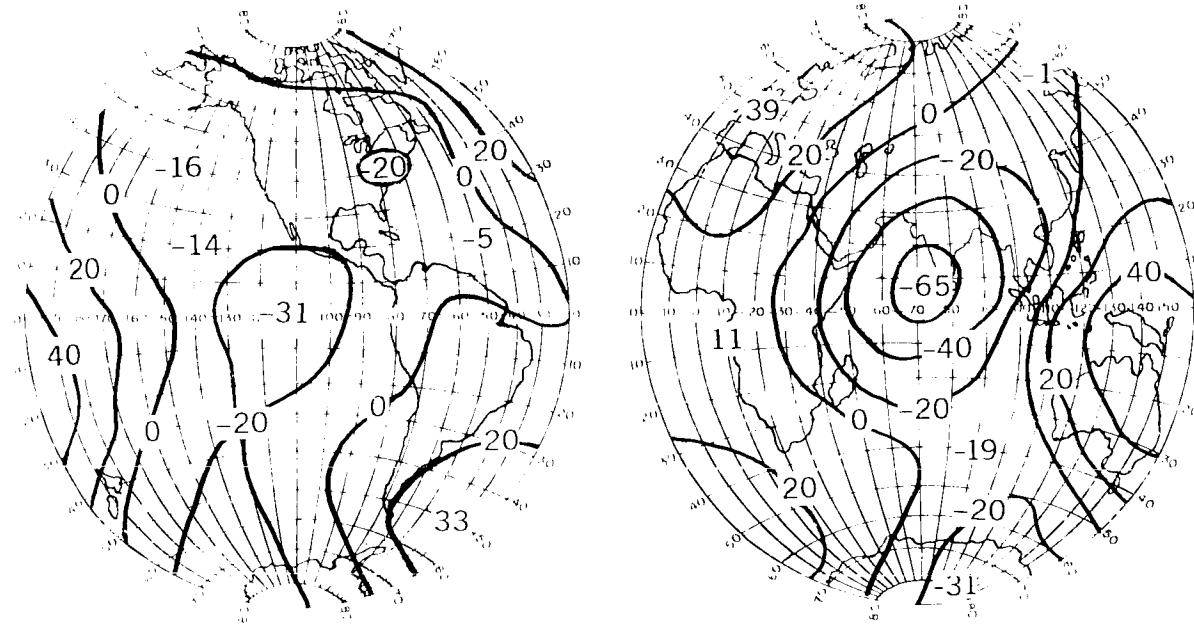


FIGURE 2.— Echo rocket geoid. Geoid heights, in meters, referred to an ellipsoid of flattening 1/298.24, determined from observations of satellite $1960\iota_2$.

TABLE 4.—*Semidaily Perturbations of Satellite Orbits*

Satellite	Element (El)	$\partial El/\partial \bar{C}_{22}$	$\partial El/\partial \bar{C}_{42}$	$\partial El/\partial \bar{C}_{62}$	$10^6 \times$comp ΔEl_c	$\partial El/\partial \bar{S}_{22}$	$\partial El/\partial \bar{S}_{42}$	$\partial El/\partial \bar{S}_{62}$	$10^6 \times$comp ΔEl_s
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$1959\alpha_1$ and	M	−2.92	0.36	5.74	−3.93	2.92	−0.36	−5.74	−3.91
1959η	i	3.62	−5.89	3.83	5.33	3.62	−5.89	3.83	−9.12
combined.	ω	1.68	12.21	−24.58	1.20	−1.68	−12.21	24.58	−7.09
	Ω	−5.54	4.22	4.88	−7.76	5.54	−4.22	−4.88	−11.10
$1960\iota_2$	M	−6.04	0.00	−0.86	−12.05	6.04	0.00	0.86	−9.86
	i	5.48	−5.08	−0.34	10.81	5.48	−5.08	−0.34	−10.76
	ω	−2.59	20.86	−2.61	−4.76	2.59	−20.86	2.61	−11.70
	Ω	−5.07	−3.26	7.66	−10.12	5.07	3.26	−7.66	−7.07

The principal sources of systematic error likely to be common to satellites $1959\alpha_1$, 1959η, and $1960\iota_2$ seem to be (1) that the magnitudes of the results will be influenced by the preassigned variances and (2) that the relative positions of tracking stations on the same geodetic datum may be appreciably in error.

For a parameter whose effects are fairly distinct in periodicities, etc., from those of other parameters, it is implausible that its preassigned variance could cause a correction that is too large or of wrong sign, but it might cause a correction that is too small. However, the variance actually used in the analyses is not the estimated squared magnitude of the correction $\sigma^2(c)$, but rather $N\sigma^2(c)$, where N is the number of orbital arcs in a set. Since N was always between 10 and 15, this seems to be no more than a mild restraint preventing occasional ill-conditioned arcs from obtaining absurdly large corrections beyond the range of linearity.

Distortion caused by the preassigned variances seems most likely to occur in separating gravitational coefficients whose principal effects are of the same period; i.e., coefficients $J_n\dot{x}$ and J_{kl} such that $m=l$ and $n-k$ is even. The most prominent set of such coefficients is J_{22}, J_{42}, and J_{62}, all of which cause semidaily variations of argument $2(\Omega-\theta)$. A way of removing some (but not all) of the influence of the preassigned covariances would be to assume that what we have determined is not the coefficients themselves but the amplitudes of semidaily variations in the orbital elements; e.g., for the $\cos 2(\Omega-\theta)$ term in the variation of the inclination

$$\Delta i = \frac{\partial i}{\partial \bar{C}_{22}}\bar{C}_{22} + \frac{\partial i}{\partial \bar{C}_{42}}\bar{C}_{42} + \frac{\partial i}{\partial \bar{C}_{62}}\bar{C}_{62} \qquad (2)$$

The semimajor axis and the eccentricity have no semidaily variation. If we omit the $1961\delta_1$ and $1961\alpha\delta_1$ results and assume that the similar $1959\alpha_1$ and 1959η orbits should be combined, we have two sets of eight equations for three unknowns. Using values $\bar{C}_{22}=1.315\times10^{-6}$, $\bar{S}_{22}=-1.473\times10^{-6}$, $\bar{C}_{42}=-0.101\times10^{-6}$, $\bar{S}_{42}=0.567\times10^{-6}$, $\bar{C}_{62}=-0.009\times10^{-6}$, $\bar{S}_{62}=-0.104\times10^{-6}$ for the combined $1959\alpha_1$ and 1959η solution (corresponding to figure 1) and using values from table 3 for $1960\iota_2$ we get the computed amplitudes of periodic perturbations in columns 6 and 10 of table 4. Using these amplitudes as the observation equation constants and solving by the rule of minimizing $\sum(d\Delta El)^2$ yields

$$\bar{C}_{22}=1.85\times10^{-6} \qquad \bar{S}_{22}=-1.75\times10^{-6}$$

$$\bar{C}_{42}=0.05\times10^{-6} \qquad \bar{S}_{42}=\ 0.34\times10^{-6}$$

$$\bar{C}_{62}=0.10\times10^{-6} \qquad \bar{S}_{62}=-0.22\times10^{-6}$$

All the coefficients are increased over the mean in table 3 except $\bar{S}_{42}$, which hints of ill conditioning. However, it looks as though only $\bar{C}_{62}$ and $\bar{S}_{62}$ might have been significantly reduced by the preassigned-variance method.

The assumption made by Kaula (1963a) that the relative positions of tracking stations on the same datum should be known through the triangulation networks with a rms error of ±20

meters or less was based on standard methods of estimating triangulation accuracy (Bomford, 1962, pp. 143–159), as is confirmed by the misclosures of large loops of triangulation: (1) 15 m in the 4000-km loop around the western Mediterranean (Whitten, 1952); (2) less than 25 m in the 10,000-km loop around the Caribbean (Fischer, 1959); and (3) within 15 m for the 10,000-km loop around the Black Sea and Caspian Sea, through Turkestan, and connecting in northwest India (Fischer, 1961).

The connections between the three northern hemisphere stations of the EASI system are closely associated with loops 1 and 3, and the connections between the three northern hemisphere stations of the Am. system are closely associated with loop 2. More in doubt are the positions of the stations in the southern hemisphere in Peru and South Africa, which depend on long single arcs of triangulation. A test run was therefore made on all the $1960\iota_2$ data, in which the stations Arequipa (in Peru) and Olifantsfontein (in South Africa) were assumed to be on separate datums. The results of this test corroborated the assumption as to triangulation accuracy; the station in Peru moved 24 m with respect to those in North America, while the station in South Africa moved 14 m with respect to those in Eurasia. The changes in the gravitational coefficients were insignificant $\bar{C}_{22}$, from 1.99 to 2.11×10^{-6}; $\bar{S}_{22}$, from -1.63 to -1.60×10^{-6}; $\bar{C}_{31}$, from 1.53 to 1.49×10^{-6}; $\bar{C}_{41}$, from -0.33 to -0.28×10^{-6}; etc.- and the maximum effect on any geoid height in figure 2 was 4 m.

There still exists the possibility of errors in the local connection of tracking stations to the triangulation systems, a matter in which better standardization of procedures is needed (Kaula, (1963b). To check this type of error for stations on the major datums we calculate the geometric geoid heights corresponding to the final positions in rectangular coordinates and then compare these heights with the gravitational geoid heights in figure 3. To estimate the size of discrepancies to be expected, we have the geoid height variance of 1076 m^2 from autocovariance analysis of gravimetry (Kaula, 1959) and a mean square height of the satellite geoid of 166 m^2, obtained from the sum of the squares of the coefficients in table 3. If the station positions and the equatorial radius were correct, the rms expected discrepancy between the geometric and gravitational geoid

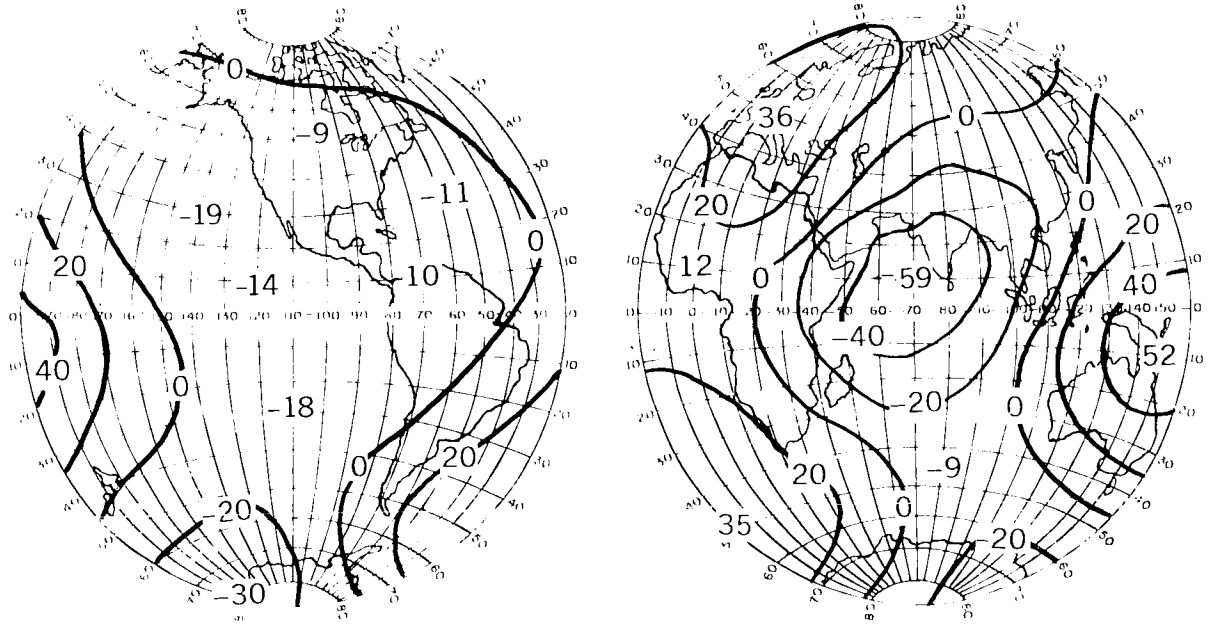


FIGURE 3.- Combined geoid. Geoid heights, in meters, referred to an ellipsoid of flattening 1/298.24, determined from observations of satellites $1959\alpha_1$, 1959η, $1960\iota_2$, $1961\delta_1$, and $1961\alpha\delta_1$.

TABLE 5.—*Comparison of Geometric and Gravitational Geoid Heights*

Station	Datum	Geometric geoid height, m	Gravitational geoid height, m	Discrepancy for 6,378,165 m radius, m	Discrepancy for 6,378,196 m radius, m
Organ Pass	Am	−4	−10	+6	−25
Arequipa		−16	−8	−8	−39
Curaçao		−25	−10	−15	−46
Jupiter		−18	−9	−9	−40
Olifantsfontein	EASI	+22	+5	+17	−14
San Fernando		+117	+26	+91	+60
Naini Tal		−17	−41	+24	−7
Shiraz		+14	−18	+32	+1
Woomera	Au	+47	+24	+23	−8
Tokyo	JKM	+54	+4	+50	+19
Villa Dolores	Ar	+104	+4	+100	+69
Maui	H	+54	−12	+66	+35

Note.—Geoid heights referred to ellipsoid of equatorial radius 6,378,165 m, flattening 1/298.24.

heights due to the inability of the satellite orbits to pick up the shorter-wave variations would be $(1076-466)^{1/2}=\pm 25$ m.

The results of the comparison are shown in table 5. Applying the mean correction of +31 m yields a mean equatorial radius of 6,378,196±11 m and a rms discrepancy of ±38 m, which implies a rms radial position error of $(382-252)^{1/2}=\pm 29$ m. Of the stations on the major datums, the 60-m discrepancy for San Fernando causes suspicion of local connection error; however, there is also a 69-m discrepancy for Villa Dolores, which was free to move to its correct position.

The agreement of the combined solution in figure 3 with astrogeodetic (Fischer, 1961) and gravimetric (Uotila, 1962) solutions is an improvement over that in Kaula (1963a), particularly in showing a more pronounced negative in the western Atlantic. The discrepancies which exist may in part be ascribed to the method of analysis of the terrestrial data, since the agreement is appreciably better with the combination of astrogeodetic, gravimetric, and satellite zonal harmonic data of Kaula (1961), especially for western Europe.

The reference flattening of 1/298.24 is used in figures 1, 2, and 3 to facilitate comparison with the results of Kaula (1961, 1963a). The flattening equivalent to the solution obtained for $\Delta\bar{C}_{20}$ is 1/298.28. The J_2 equivalent is 1082.48×10^{-6}.

In conclusion, it can be said that better explanations are needed for the systematic discrepancies indicated by tables 3 and 5. However, considering that the observations used herein depended on reflected sunlight; that they were all made more than 3 years before the minimum of solar activity; and that the orbital specifications are far from ideal, the prospects are bright for extracting more information on the gravitational field from more recent and anticipated satellites. It will be of particular interest to push the analysis to a good determination of some sixth- or eighth-degree harmonics to see whether or not they corroborate other indicators of a weak upper mantle.

ACKNOWLEDGMENTS

I am grateful to I. G. Izsak of the Smithsonian Institution Astrophysical Observatory for discussions of methods of analysis and for providing observational data in advance of publication; to Irene Fischer of the Army Map Service for discussion of triangulation and geoid accuracy; to J. Reece of Geonautics, Inc., for suggesting the

calculation of geometric geoid heights; and to W. D. Putney of Goddard Space Flight Center for help in preparing the computer runs.

REFERENCES

Bomford, G., *Geodesy*, 2d ed., 561 pp., Oxford University Press, London, 1962.

Fischer, I., A tentative world datum for geoidal heights based on the Hough ellipsoid and the Columbus geoid, *J. Geophys. Res.*, **64**, 73–84, 1959.

———, The present extent of the astro-geodetic geoid and the geodetic world datum derived from it, *Bull. Géod.*, **61**, 245–264, 1961.

Izsak, I. G., Tesseral harmonics of the geopotential, *Nature*, **199**, 137–139, 1963.

Kaula, W. M., Statistical and harmonic analysis of gravity, *J. Geophys. Res.*, **64**, 2401–2421, 1959.

———, A geoid and world geodetic system based on a combination of gravimetric, astrogeodetic, and satellite data, *J. Geophys. Res.*, **66**, 1799–1812, 1961.

———, Tesseral harmonics of the gravitational field and geodetic datum shifts derived from camera observations of satellites, *J. Geophys. Res.*, **68**, 473–484, 1963a.

———, A review of geodetic parameters, *Paper presented at the Intern. Astron. Union Symp.*, no. 21, *The System of Astronomical Constants*; to appear in *Bull. Astron.*, 1963b.

Uotila, U. A., Harmonic analysis of worldwide gravity material, *Publ. Isostatic Inst. Intern. Assoc. Geodesy*, **39**, 18 pp., 1962.

Veis, G., Haramundanis, K., Simons, L., Stern, P., and MacDonald, J. E., Catalogues of precisely reduced observations, nos. P 1 through P 7, *Smithsonian Inst. Astrophys. Obs., Res. in Space Sci., Spec. Repts.*, **82**, 1961; **85, 91, 95, 102, 104, 106**, 1962.

Whitten, C. A., Adjustment of European triangulation, *Bull. Géod.*, **24**, 187–206, 1952.

calculation of geometric geoid heights; and to W. D. Putney of Goddard Space Flight Center for help in preparing the computer runs.

REFERENCES

BOMFORD, G., *Geodesy*, 2d ed., 561 pp., Oxford University Press, London, 1962.

FISCHER, I., A tentative world datum for geoidal heights based on the Hough ellipsoid and the Columbus geoid, *J. Geophys. Res.*, **64,** 73–84, 1959.

———, The present extent of the astro-geodetic geoid and the geodetic world datum derived from it, *Bull. Géod.*, **61,** 245–264, 1961.

IZSAK, I. G., Tesseral harmonics of the geopotential, *Nature*, **199,** 137–139, 1963.

KAULA, W. M., Statistical and harmonic analysis of gravity, *J. Geophys. Res.*, **64,** 2401–2421, 1959.

———, A geoid and world geodetic system based on a combination of gravimetric, astrogeodetic, and satellite data, *J. Geophys. Res.*, **66,** 1799–1812, 1961.

———, Tesseral harmonics of the gravitational field and geodetic datum shifts derived from camera observations of satellites, *J. Geophys. Res.*, **68,** 473–484, 1963a.

———, A review of geodetic parameters, *Paper presented at the Intern. Astron. Union Symp.*, no. 21, *The System of Astronomical Constants*; to appear in *Bull. Astron.*, 1963b.

UOTILA, U. A., Harmonic analysis of worldwide gravity material, *Publ. Isostatic Inst. Intern. Assoc. Geodesy*, **39,** 18 pp., 1962.

VEIS, G., HARAMUNDANIS, K., SIMONS, L., STERN, P., and MACDONALD, J. E., Catalogues of precisely reduced observations, nos. P–1 through P–7, *Smithsonian Inst. Astrophys. Obs., Res. in Space Sci., Spec. Repts.*, **82,** 1961; **85, 91, 95, 102, 104, 106,** 1962.

WHITTEN, C. A., Adjustment of European triangulation, *Bull. Geod.*, **24,** 187–206, 1952.

A REVIEW OF GEODETIC PARAMETERS

WILLIAM M. KAULA

Goddard Space Flight Center

It is recommended that the parametric values which are currently most used in orbital computation be adopted as provisional standards, rather than those which may be the best available, because the "most used" values differ only slightly from the "best" values and further improvements in the values are expected within the next 4 years. Some of these values are:

$GM_\oplus = 3.986032 \times 10^{20}$ cm^3 sec^{-2},
$J_2 = 1082.30 \times 10^{-6}$,
$J_3 = -2.3 \times 10^{-6}$,
$J_4 = -1.8 \times 10\sigma^6$,
$a_e = 6{,}378{,}165.0$ m.

With parameters such as the foregoing the most serious geodetic errors affecting astronomy are tracking station positions. Standard methods of describing and transforming positions are suggested.

INTRODUCTION

This review recommends which geodetic parameters should be adopted as standard, the manner in which the parameters should be expressed, and the values which should be adopted. In making these recommendations, current practice, available determinations, and anticipated improvements will be considered.

GRAVITATIONAL PARAMETERS

For the notation of the earth potential, recommendations have already been made by Commission 7 on Celestial Mechanics, of the International Astronomical Union[1]:

$$U = \frac{\mu}{r}\left[1 + \sum_{n=1}^{\infty}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^n P_n^m(\sin\beta)\left(C_{n,m}\cos m\lambda + S_{n,m}\sin m\lambda\right)\right] \tag{1}$$

where $\mu = GM_\oplus$, r is the distance from the center of the earth, R is the mean equatorial radius of the earth, P_n^m is the associated Legendre polynomial, β is the latitude, and λ is the longitude. Alternative notations recommended for the gravitational coefficients are

$$J_n = -C_{n,0}\ , \tag{2}$$

and

$$(A_{n,m}, B_{n,m}) = \left[\frac{(n+m)!}{(n-m)!}\right]^{1/2}(C_{n,m}, S_{n,m})\ . \tag{3}$$

These two additions are suggested:

1. Define

$$(\bar{C}_{n,m}, \bar{S}_{n,m}) = \left[\frac{(n+m)!}{(n-m)!\,(2n+1)\,(2-\delta_m^0)}\right]^{1/2} (C_{n,m}, S_{n,m}), \tag{4}$$

where the Dirac delta $\delta_m{}^0$ is 1 for $m=0$ and 0 for $m\neq 0$. The $\bar{C}_{n,m}$, $\bar{S}_{n,m}$ are coefficients of harmonics which have a mean square amplitude of 1 for all values of n and m.

2. Define the mean equatorial radius more precisely as the equatorial radius of the mean earth ellipsoid; i.e., the ellipsoid of revolution which best fits the geoid. This definition is consistent with geodetic practice and involves the equatorial radius with only two of the set of orthogonal parameters defining the radius vector of the geoid—the zeroth and second degree zonal harmonics. (The more literal definition of the mean equatorial radius as the radius of the circle which best fits an equatorial section through the geoid would connect the radius to the infinite set of even degree zonal harmonics.) An alternative possibility for the equatorial radius in equation 1 is the mean radius of the entire earth which, since it differs by a factor of 10^{-3}, would affect the value of J_2. The mean radius seems slightly preferable aesthetically, but current practice overwhelmingly favors the equatorial radius; a perusal of some papers on close satellite dynamics and orbit analysis found ten workers using the equatorial radius but none using the mean radius (in addition, five theoreticians did not define their radius).

To be consistent with the connection of equatorial radius to the mean earth ellipsoid, it is recommended that the following be the relationships between the astronomical parameters $\mu=GM_\oplus$ and $J_2=-C_{2,0}$ and the geodetic parameters $R=a_e$, the equatorial radius; γ_e, the equatorial gravity; f, the flattening; and ω, the rate of the earth's rotation with respect to inertial space:[2,3,4]

$$GM_\oplus = a_e{}^2\gamma_e\left[1 + \frac{3}{2}m - f - \frac{15}{14}mf - \frac{1}{294}mf^2 - O(f^4)\right] \tag{5}$$

$$J_2 = \frac{2}{3}f\left(1 - \frac{1}{2}f\right) - \frac{1}{3}m\left[1 - \frac{3}{2}m - \frac{2}{7}f + \frac{9}{4}m^2 + \frac{11}{49}f^2 + O(f^3)\right] \tag{6}$$

where

$$m = \frac{\omega^2 a_e}{\gamma_e}. \tag{7}$$

The values of $GM_\oplus$ and J_2 which are probably the most extensively used at orbit computation centers in the United States are:[5,6,7]

$$\left.\begin{aligned} GM_\oplus &= 3.986032 \pm 0.000030 \times 10^{20}\ \text{cm}^3\ \text{sec}^{-2}, \\ J_2 &= 1082.30 \times 10^{-6} \end{aligned}\right\} \tag{8}$$

In the alternative notation of Herrick, Baker, and Hilton[8] and Makemson, Baker, and Westrom[9]:

$$k_e = (GM_\oplus)^{1/2} = 0.019965049\ \text{megameter}^{3/2}\ \text{sec}^{-1}. \tag{9}$$

The values of $GM^{\oplus}$ and J_2 in equation 8 are consistent with these values for the geodetic parameters:

$$\left.\begin{aligned} a_e &= 6{,}378{,}165.0 \pm 25.0 \text{ meters}, \\ \gamma_e &= 978.0300 \pm 0.012 \text{ cm sec}^{-2}, \\ f &= 1/298.30, \\ \omega &= 0.729211585 \times 10^{-4} \text{ sec}^{-1}. \end{aligned}\right\} \quad (10)$$

The value for a_e is a compromise between the solutions of Fischer,[12] and Kaula,[11] and other values which are unpublished. The γ_e value differs from that of the International Formula and the Potsdam System (978.0490 cm sec^{-2}) in three ways:

1. Correction to Potsdam System absolute g (reference 12) $= -0.0128 \pm 0.0003$;
2. Change of flattening from 1/297 to 1/298.3 $= -0.0051$;
3. Change of mean gravity over the earth's surface (reference 11) $= -0.0005 \pm 0.0012$.

The correction to absolute g is a provisional value and has not been adopted by the International Union of Geodesy and Geophysics; an improved value should be forthcoming within the next few years from several determinations in progress.[13] The correction to mean gravity is negative, mainly because correlation between gravity and topography was used to estimate anomalies for the areas without observations, which are predominantly oceans. Solutions by Uotila which fit observed gravimetry and do not use correlation with topography give positive corrections ranging from +0.0004 to +0.0019 cm sec^{-2} (reference 14). Rather slow improvement is expected; problems in observing gravity at sea are not entirely solved.[15,16] Some improvement may also come from using the better statistical techniques which larger capacity computers permit.

The value of $GM_{\oplus}$ may also be obtained through the modified Kepler equation by using the radar mean distance of the moon A and the moon's mean motion n:

$$GM_{\oplus} = \frac{n^2(1+\beta)^3}{1+\frac{\mu_M}{\mu_E}} A^3, \quad (11)$$

where β is the solar perturbation of the mean semimajor axis and μ_M/μ_E is the ratio of the moon's mass to the earth's mass, equal to the lunar inequality.[17] The most recently published value for A is 384,402.1±.2 km.[18] As pointed out] by Fischer, this value should perhaps be corrected because it is dependent on an excessively rounded-off lunar radius of 1740 km.[19] The mean radius of the lunar limb is 1737.85±0.07 km. Geometrical determinations of the radius toward the earth vary considerably; Baldwin's conclusion[20] leads to 1740.05 km, whereas Schrutka-Rechtenstamm[21] concludes that the bulge is too small to be determined. However, we are not interested in just the long axis of a best-fitting triaxial ellipsoid, but rather in the mean radius of the area contributing to the leading edge of the radar return pulse, which would fall within the ±7° area of libration. Contour maps of the moon (reference 22, for example) indicate that the average radius of this ±7° area could differ by as musch as 2 km from the best-fitting ellipsoid. If the lunar surface is assumed to be an equipotential surface, then using the moments of inertia obtained from the physical libration yields 1738.57 km as the radius toward the earth. Letting $A = 384{,}400.5 \pm 1.2$ km, $\beta = 0.0090678$, $n = 2.6616997 \times 10^{-6}$ sec^{-1} (reference 23), and $\mu_M/\mu_E = 1/(81.375 \pm 0.026)$ (reference 24) gives

$$GM_{\oplus} = 3.986094 \pm 0.00004 \times 10^{20} \text{ cm}^3 \text{ sec}^{-2}. \quad (12)$$

Using the $\mu_M/\mu_E = 1/81.219$ of Delano[25] reduces $GM_{\oplus}$ to 3.986001×10^{20} cm^3 sec^{-2}, so the difference from solutions based on terrestrial data seems largely explicable as an error in the lunar inequality. The larger computers of today permit the application of more elaborate statistical techniques than it was possible to apply in 1950 (the year Delano and Rabe published their work).

However, since the stellar positions are a major suspect for systematic error, it seems premature to reanalyze the Eros observations before the revised reference star systems are available.[26] Meanwhile, improved determination of the lunar inequality may be obtained from radio tracking of space probes such as Mariner II (1962 $\alpha p1$). Also, since spacecraft have been launched into high, nearly circular orbits such as those of Midas III (1961σ) and Midas IV (1961 $\alpha\delta$), it may be worthwhile to try to determine $GM_\oplus$ from close satellite orbits.

In addition to $GM_\oplus$ and J_2, standard orbit computation programs usually incorporate J_3 and J_4. The values which are probably most common at U.S. computation centers are[6]:

$$\left.\begin{aligned} J_3 &= -2.3 \times 10^{-6}, \\ J_4 &= -1.8 \times 10^{-6}. \end{aligned}\right\} \qquad (13)$$

At present the best values of the zonal harmonics are undoubtedly those of Kozai[7]:

$$\left.\begin{aligned} J_2 &= 1082.48 \pm 0.06 \times 10^{-6}, & J_3 &= -2.562 \pm 0.012 \times 10^{-6}, \\ J_4 &= -1.84 \pm 0.08 \times 10^{-6}, & J_5 &= -0.064 \pm 0.019 \times 10^{-6}, \\ J_6 &= 0.39 \pm 0.12 \times 10^{-6}, & J_7 &= -0.470 \pm 0.021 \times 10^{-6}, \\ J_8 &= -0.02 \pm 0.02 \times 10^{-6}, & J_9 &= 0.117 \pm 0.025 \times 10^{-6}. \end{aligned}\right\} \qquad (14)$$

Note that the J_2, J_3, and J_4 now used, given in equations 8 and 13, each differ from Kozai's improved values by less than 0.3×10^{-6}; and that the coefficients J_5 and higher are all very small in absolute magnitude. Therefore, it does not seem worthwhile to adopt values, other than those already in general use, before 1966 or 1967, when analysis of geodetic satellite orbits observed during the International Year of the Quiet Sun will be completed.

Most of the current close satellite orbit analyses for geodetic purposes seek tesseral harmonic perturbations. In view of the smallness of these perturbations, it does not seem appropriate to adopt standardized values for the tesseral harmonics $C_{n,m}$, $S_{n,m}$. The one exception might be $C_{2,2}$, $S_{2,2}$, for which an upper limit would be useful because of its effect on supplemental energy requirements for 24-hr orbits. The most recent, unpublished determinations of Izsak, Kaula, Kozai, and Newton range from 0.9×10^{-6} to 1.8×10^{-6} in amplitude $(C_{2,2}^2+S_{2,2}^2)^{1/2}$ and from 8° to 25° W in the direction of the principal axis $[\frac{1}{2}\tan^{-1}(S_{2,2}/C_{2,2})]$.

GEOMETRICAL PARAMETERS

As shown by analyses involving large systems of observations,[10,11,19] the equatorial radius is a derived, rather than a fundamental, quantity: accurate knowledge of the radius is not necessary to obtain other parameters, such as the lunar distance, geoid undulations, or datum positions by fitting of the astro-geodetic to the gravimetric geoid. However, for astronomical purposes, it is desirable to have a reference ellipsoid correct within ±50 m in order to obtain reasonably correct positions of isolated tracking stations from astronomic latitude and longitude. Also it is convenient to have a unit of length approximating the earth's radius for use in the potential formula (equation 1) and for use as a base line to compare or combine parallax observations. For these astronomical purposes, the value of 6,378,165.0 m

TABLE 1.—*Corrections to u, v, w from Reference 11 (meters).*

Datum Shift	Δu	Δv	Δw
WGS-NAD	−23 ± 26	+142 ± 22	+196 ± 22
WGS-ED	−57 ± 23	−37 ± 29	−96 ± 23
WGS-TD	−89 ± 40	+551 ± 53	+710 ± 40

given in equation 10 should be entirely adequate. Marked improvement is not expected for about 5 years, by which time satellite observations should contribute significantly to the strengthening of triangulation systems and to the interconnection of geodetic datums.

By far the most annoying problems in the astronomical application of geodetic data pertain to tracking station positions. Errors in the adopted values of station positions, in conjunction with drag and nonuniform distribution of observations, prevent accurate determination of tesseral harmonics and are even believed to be a major cause of discrepancies in space probe trajectories.[27] These station position errors are due to both inadequate data and mistaken treatment of data; in descending order of reprehensibility they include:

1. Weak, erroneous, or nonexistent connection of tracking stations to local geodetic control (this includes the moving of antennas by stations without informing the computing center);
2. Failure to state the datum or ellispoid to which tracking station positions refer;
3. Use of obsolete or erroneous standard datum and ellipsoid;
4. An incomplete or ambiguous statement about how datum or ellipsoid transformations were made;
5. Failure to provide for geoid-ellipsoid difference in calculating heights;
6. Neglecting systematic error due to incorrect observation (for example, no Laplace stations) or incorrect adjustment (for example, arbitrary scale changes or rotations) of geodetic control connecting tracking stations more than, say, 1000 km apart;
7. Actual observational error of position.

In view of the number of geodetic datums and corrections thereto, they do not seem to be appropriate parameters to be adopted as standard by an international organization, except possibly for the large continental triangulation systems. The corrections to coordinates u, v, w with positive axes directed respectively toward latitude and longitude (0°, 0°), (0°, 90° E), (90° N) obtained in the world geodetic system solution of Kaula[11] are listed in table 1, where NAD, ED, and TD refer to the North American, European, and Tokyo datums, respectively. The uncertainties in this table are based on estimates of the errors due to interpolation and representation in the astrogeodetic and gravimetric geoids, and are probably a fair measure of item 7 on the above list, but may neglect significant effects falling under item 6. The relationships of the rectangular coordinates u, v, w to the geodetic latitude ϕ, longitude λ, and elevation h, referred to an ellipsoid of parameters a_e and f, are:

$$\left.\begin{aligned} u &= (\nu + h)\cos\phi\cos\lambda\,, \\ v &= (\nu + h)\cos\phi\sin\lambda\,, \\ w &= \left[(1-e^2)\,\nu + h\right]\sin\phi\,, \end{aligned}\right\} \qquad (15)$$

where $\nu = a_e/(1 - e^2\sin^2\phi)^{1/2}$ and $e^2 = 2f - f^2$.

To help minimize the number of unnecessary errors in categories 1 through 5 on the above list, it is suggested that organizations be urged to publish the following information pertaining to each tracking station for which they publish any precise observations of artificial satellites or probes, or orbital data based thereon:

1. The names and coordinates of local geodetic control points, both horizontal and vertical, to which the tracking station is connected;
2. The geodetic datum and ellipsoid to which the horizontal coordinates refer;
3. The organization which established the local geodetic control points;
4. The manner in which the horizontal and vertical survey connections were made from the local control points to the tracking station;
5. The date of the survey connection and a description of the termination point of the survey;
6. The geodetic (ϕ, λ, h) and rectangular (u, v, w) coordinates of the station referred to the local deodetic datum;
7. A statement of the geoid height, if any, estimated for the station and the basis for the estimate;
8. If the tracking station position has been shifted for the purpose of referring observations (direction cosines or altitude and azimuth) or calculating orbits, the geodetic and rectangular coordinates after the shift and the ellipsoid to which the new coordinates refer.

Every item on this list is an action which must be accomplished for any tracking station, but thus far the Smithsonian Astrophysical Institute Baker-Nunn camera network is the only one for which even part of the list has been published.[28] It is symptomatic of the difficulties which occur that, since this publication, the coordinates for at least 4 of the 12 Baker-Nunn cameras have been found to be in error by 20 m or more. These geometrical details of tracking station position are rather uninteresting, but they must be examined carefully and determined correctly if the full potentialities of modern tracking techniques are to be realized.

REFERENCES

1. Hagihara, Y., "Recommendations on Notation of the Earth Potential," *Astronom. J.*, **67**(1): 108, February 1962.
2. Lambert, W. D., "The Gravity Field of an Ellipsoid of Revolution as a Level Surface," *Annales Academiae Scientiarum Fennicae*, Ser. A–III, No. 57, 1961; Reprinted in Ohio State Univ., Inst. of Geodesy, Photogrammetry and Cartography, Publ. No. 14, 1961.
3. Cook, A. H., "The External Gravity Field of a Rotating Spheroid to the Order of e^3," *Geophys. J.*, **2**(2): 199–214, September 1959.
4. Hirvonen, R. A., "New Theory of the Gravimetric Geodesy," *Annales Academiae Scientiarum Fennicae*, Ser. A–III, No. 56, 1960; Reprinted in Ohio State Univ., Inst. of Geodesy, Photogrammetry and Cartography, Publ. No. 9, 1960.
5. Kaula, W. M., "Tesseral Harmonics of the Gravitational Field and Geodetic Datum Shifts Derived from Cambera Observations of Satellites," *J. Geophys. Res.*, **68**(2): 473–484, Jan. 15, 1963.
6. Clarke, V. C., "Constants and Related Data Used in Trajectory Calculations at the Jet Propulsion Laboratory," Calif. Inst. Tech., Jet Propulsion Lab., Tech. Rept., **32**, 273, May 1, 1962.
7. Kozai, Y., "Numerical Results from Orbits," Smithsonian Inst., Astrophys. Observ. Spec. Rept., No. 101, July 31, 1962.
8. Herrick, S., Baker, R. M. L., Jr., and Hilton, C. G., "Gravitational and Related Constants for Accurate Space Navigation," in: *Proc. 8th Internat. Astronaut. Cong., Barcelona*, 1957, ed. by F. Hecht, Vienna: Springer-Verlag, 1958, pp. 197–235.
9. Makemson, M. W., Baker, R. M. L., Jr., and Westrom, G. B., "Analysis and Standardization of Astrodynamic Constants," *J. Astronaut. Sci.*, **8**(1): 1–13, Spring 1961.
10. Fischer, I., "An Astrogeodetic World Datum from Geoidal Heights Based on the Flattening $f = 1/298.3$," *J. Geophys. Res.*, **65**(7): 2067–2076, July 1960.
11. Kaula, W. M., "A Geoid and World Geodetic System Based on a Combination of Gravimetric, Astrogeodetic, and Satellite Data," *J. Geophys. Res.*, **66**(6): 1799–1811, June 1961.
12. Rice, D. A., "Compte rendu des réunions de la Section IV—Gravimetrié," *Bulletin Géodésiqué* No. 60, June 1, 1961, p. 109.
13. Cook, A. H., "Report on Absolute Measurements of Gravity," *Bulletin Géodésique* No. 60, 131–139, June 1, 1961.
14. Uotila, U. A., "Corrections to Gravity Formula from Direct Observations and Anomalies Expressed in Lower Degree Spherical Harmonics," Ohio State Univ., Inst. of Geodesy, Photogrammetry and Cartography, Publ. No. 23, 1962.
15. Allan, T. D., Dehlinger, P., et al., "Comparison of Graf-Askania and LaCoste-Romberg Surface-Ship Gravity Meters," *J. Geophys. Res.*, **67**(13): 5157–5162, December 1962.
16. Harrison, J. C., "The Measurement of Gravity," *Proc. IRE*, **50**(11): 2302–2312, November 1962.
17. O'Keefe, J. A., Eckels, A., and Squires, R. K., "The Gravitational Field of the Earth," *Astronom. J.*, **64**(7): 245–253, September 1959.

18. BRUTON, R. H., CRAIG, K. J., and YAPLEE, B. S., "The Radius of the Earth and the Parallax of the Moon from Radar Range Measurements on the Moon," *Astronom. J.*, **64**(8): 325, October 1959 (Abstract).
19. FISCHER, I., "Parallax of the Moon in Terms of a World Geodetic System," *Astronom. J.*, **67**(6): 373–378, August 1962.
20. BALDWIN, R. B., "The Face of the Moon," Chicago: Univ. of Chicago Press, 1949.
21. SCHRUTKA-RECHTENSTAMM, G., "Neureduktion der 150 Mondpunkte der Breslauer Messengun von J. Franz," *Sitzungsberichte der Österreichischen Akademie der Wissenschaften, Mathematisch-Naturwissenschaftliche Klasse*, Abt. II, **167,** 71–123, 1958.
22. BALDWIN, R. B., "A Lunar Contour Map," *Sky and Telescope*, **21**(2): 84–85, February 1961.
23. BROWN, E. W., "Theory of the Motion of the Moon, Part 4," *Mem. Roy. Astronom. Soc.*, **57,** 51–145, 1908.
24. RABE, E., "Derivation of Fundamental Astronomical Constants from the Observations of Eros During 1926–1945," *Astronom. J.*, **55**(4): 112–126, May 1950.
25. DELANO, E., "The Lunar Equation from Observations of Eros, 1930–1931," *Astronom. J.*, **55**(5): 129–132, August 1950.
26. SCOTT, F. P., "Status of the International Reference-Star Programs," *Astronom. J.*, **67**(10): 690–695, December 1962.
27. HAMILTON, T. W., "Applications of Celestial Mechanics to Spacecraft Flight," in: *Proc. of the NASA-University Conf. on the Sci. and Tech. of Space Exploration, Chicago, November* 1962, NASA SP–11, December 1962, vol. 1, pp. 253–260.
28. VEIS, G., "The Positions of the Baker-Nunn Camera Stations," Smithsonian Inst., Astrophys. Observ. Spec. Rept. No. 59, Mar. 3, 1961.

SATELLITE ORBIT ANALYSES FOR GEODETIC PURPOSES

W. M. KAULA

Goddard Space Flight Center

The difficulties in obtaining accurate determinations of tracking station positions and longitudinal variations of the earth's gravitational field are not problems of mechanics, but rather of data analysis: nonuniform distribution of observations, inadequacy of the mathematical model of the atmosphere, and similarity of effects of different parameters on the same orbit.

Because the characteristic periocities of the geodetic effects fall in a different part of the spectrum of orbital variations from drag effects, the geodetic effects should, in theory, be determinable even though appreciably smaller. Given a statistical model for drag effects, a generalization of linear autoregression theory can be developed, which indicates, however, that in the computation arrays of dimensions on the order of the number of observations must be manipulated. Hence either the number of observations treated must be limited or the rigorous treatment modified. The second alternative has been applied in analyzing Baker–Nunn camera observations of satellites 1959α-1, 1959η-1, 1960ι-2, in which various empirical methods are applied in place of allowing for the correlation between different observations. Some preliminary results obtained from 1960ι-2 are presented.

I. INTRODUCTION

The geometrical[1,2] and gravitational[2,3] problems in analyzing close satellite orbits for geodetic parameters beyond the oblateness—station positions and the higher harmonics of the gravitational field—are relatively straight forward, since the effects involved are small enough to be treated as first-order perturbations. Since there are numerous independent parameters, of which several have similar effects on a given orbit, two factors—the nonuniform distribution of observations and the inadequacy of the atmospheric model—require attention to some statistical problems to attain fullest geodetic exploitation of any accurate satellite observations. This paper has two main parts: (1) development of the appropriate statistical theory and (2) numerical results obtained thus far from simpler approximate methods.

II. THEORY OF NONUNIFORMLY DISTRIBUTED OBSERVATIONS

We start from the generalized linear autoregression of prediction theory,[4,5] which states that, given $x(t)$ over and interval $A<t<B$, the optimum prediction (in a least-squares sense) $E\{x(z)\}$, for $z>B$, can be expressed as

$$E\{x(z)\}=\int_A^B w(z,t)x(t)\,dt, \tag{1}$$

where $w(z,t)$ is the solution of

$$K(z,s)=\int_A^B w(z,t)K(t,s)\,dt, \quad A<s<B. \tag{2}$$

$K(u,v)$ is the covariance between $x(u)$ and $x(v)$.

The modifications of (1) and (2) appropriate to our problem:

(a) The variable x is a 6-dimensional vector—position and momentum of a satellite—so make the replacements:

$$\begin{aligned} x(z)&\rightarrow x_f(z),\\ w(z,t)&\rightarrow w_{fh}(z,t),\\ K(z,s)&\rightarrow K_{fg}(z,s). \end{aligned} \tag{3}$$

(b) Observations are not continuous over the interval A, B, but are in several disconnected

bits, so make the replacements:

$$\int_A^B \cdots dt \rightarrow \sum_k \int_{a_k}^{b_k} \cdots dt = \int_A^B I(t) \cdots dt, \quad (4)$$

where minimum $a_k \geq A$, maximum $b_k \leq B$, $I(t)=1$, $a_k < t < b_k$, and $I(t)=0$, $b_k < t < a_{k+1}$ for any k.

(c) The $x_h(t)$ are not observed, but rather a linear transform $y_j(t)$ of lesser dimension (e.g., 2 for photos, 1 for range or range-rate), so make the replacements

$$x_h(t) \rightarrow y_j(t),$$

$$w_{fh}(z, t) \rightarrow w_{fh}(z, t)\frac{\partial y_j(t)}{\partial x_h(t)} = v_{fj}(z, t),$$

$$K_{fg}(z, s) \rightarrow K_{fg}(z, s)\frac{\partial y_c(s)}{\partial x_g(s)} = M_{fc}(z, s), \quad (5)$$

$$K_{hg}(t, s) \rightarrow \frac{\partial y_j(t)}{\partial x_h(x)} K_{hg}(t, s)\frac{\partial y_c(s)}{\partial x_g(s)} = L_{jc}(t, s).$$

In (5), and hereafter, the rule of summation over repeated subscripts in a product applies.

(d) The observations $y_j(t)$ may be affected by errors in certain parameters (e.g., datum position), so make the replacement

$$y_j(t) \rightarrow \hat{y}_j(t) - [\partial y_j(t)/\partial q_m] E\{q_m\}, \quad (6)$$

where $E\{q_m\}$ is the correction to q_m.

(e) We may not be interested in the $x_f(z)$, but rather certain parameters p_l of which they are functions (e.g., departures from secularly changing Keplerian elements are functions of parameters of the gravitational field and the atmospheric model), so make the replacement

$$E\{x_f(z)\} \rightarrow [\partial x_f(z)/\partial p_l] E\{p_l\}. \quad (7)$$

Putting (3) through (7) in (1) and (2), we get

$$\frac{\partial x_f(z)}{\partial p_l} E\{p_l\}$$

$$= \int_A^B I(t) v_{fj}(z, t) \left[\hat{y}_j(t) - \frac{\partial y_j(t)}{\partial q_m} E\{q_m\} \right] dt. \quad (8)$$

$$M_{fc}(z, s) = \int_A^B I(t) v_{fj}(z, t) L_{jc}(t, s)\, dt. \quad (9)$$

Over a finite interval of time, $x_f(z)$ which are departures from a model of secularly changing Keplerian elements (or a model taking into account any major periodic variations, such as those due to J_2, as well) can be represented by a Fourier expansion

$$x_f(z) = \Re \sum_{n=0}^{\infty} \alpha_{fn} \exp[i\lambda_n z]. \quad (10)$$

The α_{fn} are complex; $\Re$ denotes the real part, and i is $(-1)^{1/2}$. In any practical application, we are concerned only with the finite interval $A < z < B$, so

$$\lambda_n = 2\pi n/(B-A). \quad (11)$$

Also, in any practical case, a finite upper limit can be taken for n; for satellite orbits, certainly one such that $1/\lambda_n > 20$ minutes.

Apply the Fourier transformation

$$\int_A^B \cdots \exp(i\lambda_n z)\, dz$$

to (8) and (9):

$$\frac{\partial \alpha_{fn}}{\partial p_l} E\{p_l\}$$

$$= \int_A^B I(t) \Omega_{fjn}(t) \left[\hat{y}_j(t) - \frac{\partial y_j(t)}{\partial q_m} E\{q_m\} \right] dt. \quad (12)$$

$$\rho_{fcn}(s) = \int_A^B I(t) \Omega_{fjn}(t) L_{jc}(t, s)\, dt, \quad (13)$$

where $\rho_{fcn}(s)$ and $\Omega_{fjn}(t)$ are the transformation of $M_{fc}(z, s)$ and $v_{fj}(z, t)$, respectively.

The advantage of the Fourier representation in (12) is apparent. For the effect of any gravitational harmonic, which has a discrete spectrum, most of the $\partial \alpha_{fn}/\partial p_l$ are zero, while in the equations where they are non-zero, the $\partial \alpha_{fn}/\partial p_l$ due to drag effects are negligibly small.

To solve (13) for the $\Omega_{fjn}(t)$ to use in (12), there are needed statistical models for the gravitational and drag effects on the orbit. Given these models, the $L_{jc}(t, s)$ can be constructed from the $K_{hg}(t, s)$ by (5), while to obtain the $\rho_{fcn}(s)$ the Fourier transform with respect to z must be applied to $K_{fg}(z, s)$ before applying the post-multiplication $\partial y_c(s)/\partial x_g(s)$.

The statistical model for the gravitational part

TABLE 1.—*Coefficients for Normalized Spherical Harmonics* Obtained From Satellite 1960ι-2 ($h_\pi = 1500$ km, $i = 47°$, $e = 0.01$, $A/m = 0.2$ cm²/gm) Sept.–Dec. 1960. Five Orbital Arcs of 23 Days Each. Multiply by Scaling Factor 10^{-6}*

Number of observations	$\bar{C}_{21}$	$\bar{S}_{21}$	$\bar{C}_{22}$	$\bar{S}_{22}$	$\bar{C}_{1}$	$\bar{S}_{1}$	$\bar{C}_{41}$	$\bar{S}_{41}$
151	0.17	−0.13	1.72	−1.46	1.82	0.00	0.09	0.16
179	−0.05	0.39	1.10	−2.91	1.43	−1.44	−0.68	1.13
126	−0.13	−0.01	2.37	−1.57	0.34	0.12	−0.71	0.09
113	−0.03	0.12	2.55	−0.96	1.76	0.03	0.03	0.64
110	0.26	0.25	1.44	−1.67	2.26	−0.39	−0.35	.420
679	0.04	0.13	1.77	−1.81	1.50	−0.41	−0.34	0.52

* Functions Y_{nm} such that $\int_{\text{sphere}} Y_{nm}^2 \, d\sigma = 4\pi$.

of K_{fg} is obtainable from the autocovariance analysis of terrestrial gravimetry,[6] the orders-of-magnitude of which have been confirmed by the zonal harmonics thus far obtained from satellite orbits.[7]

For the drag part of K_{fg} the appropriate data are less sure, mainly due to the considerable variation from one year to another and from one altitude to another. Such statistical analyses of drag as have so far been carried out[8,9] have been on a limited variety of orbital specifications.

III. NUMERICAL ANALYSIS OF ACTUAL ORBITS

The aforedescribed theoretical development has not been applied in numerical computations because it has not yet been worked out to what extent there can be modified the computational manipulation of arrays of dimension comparable to the number of observations, as is implied by Eq. (13). Instead, it has been attempted to avoid allowing for covariance between different observations by various devices, of more-or-less intuitive justification, as described in references 9, 10:

1. Inclusion in the reference model to the maximum extent possible of the various longterm effects: zonal harmonics, luni-solar perturbations, radiation pressure, and drag by physical or empirical models.
2. Weighting of observations inversely as their density with respect to the angles critical for the determination of the parameters of interest.
3. Use of the covariance matrix of parameters as a means of combining results of different satellites, or of satellite with terrestrial results.
4. Increasing the variance of observations and

TABLE 2.—*Datum Shift for Woomera, Australia, Obtained From Satellite 1960ι-2 Sept.–Dec. 1960. Units of Earth Radii. Multiply by Scaling Factor 10^{-6}*

Number of observations from Woomera	dU	dV	dW
30	−17.1	−16.3	5.2
25	3.3	39.6	3.6
9	2.4	−9.9	0.1
20	−4.8	−19.4	4.0
26	−18.3	−,2.2	1.8
110	−8.9	−2.7	3.4

U, V, W axes toward lat., long. of 0°, 0°; 0°, 90° E; 90° N, respectively.

their timing over that which can be reasonably ascribable to the observations and timing themselves.

5. Use of arbitrary polynomials.

From the viewpoint of the more rigorous theory discussed in sec. II, devices 1 and 5 attempt to reduce the segment of the spectrum which must be considered; device 2 is an approximation of what should be obtained for $\Omega_{fjn}(t)$; device 2 is an approximation of what should be obtained for $\Omega_{fjn}(t)$; device 3 extends the set of observations used beyond those described by $I(t)$, $A<t<B$, in the modified form of the solution obtained by the added observations; and device 4 in effect arbitrarily shifts the higher frequency (of the order 0.1 to 1.0 cycle/day) part of the drag spectrum to a band of much higher frequency than that of the observations.

These methods are being applied to Baker–Nunn camera observations of satellites 1959α-1, 1959η, and recently, 1960ι-2. These analyses are being made for those high frequency (daily, semi-daily, etc.) effects which should be expected to cause perturbations of more than $\pm 10^{-5}$ or ± 70 m: the harmonics of the gravity field up to n, $m=4$, 2, and the station positions for the cameras located in Australia, Japan, Argentina, and Hawaii. The atmospheric model of Jacchia[11] is applied to the two Vanguard orbits (1959α-1 and 1959η), but none is used for the high, dense Echo I rocket case (1962ι-2). Orbital arcs of about 20 days are used, and the arbitrarily polynomials included go to t^4 in the mean anomaly plus secular terms for the inclination and eccentricity. To absorb some of the high frequency drag effect, the timing standard deviation is assumed to be ± 0.007 sec. The angle-weighting used is with respect to the angle (node-GST), integral multiples of which appear as the arguments of the principal tesseral harmonic effects. The observational residuals obtained from the solution for each satellite are on the order of $\pm 10''$ in direction and $\pm 0.03^s$ time. The tests of agreement between solutions from different orbital arcs and of a negligibly small answer for the "impossible" harmonic J_{21} have as yet been satisfied only by the analysis of 1960ι-2, for which preliminary results are shown for some gravitational coefficients in table 1 and for the position of Woomera, Australia (which had the most observations) in table 2. Since these arcs happen to fall across the events of October and November 1960, appreciable improvement is expected.

REFERENCES

1. Veis, G.: Geodetic uses of satellites. Smithsonian Contrib. to Astrophys. **3,** 95–161 (1960).
2. Kaula, W. M.: Analysis of gravitational and geometric aspects of geodetic utilization of satellites. *Roy. Astr. Soc. Geophys. J.* **5,** 104–133 (1961).
3. Groves, G. V.: Motion of a satellite in the earth's gravitational field. *Proc. Roy. Astr. Soc.* **254A,** 48–65 (1960).
4. Bartlett, M. S.: "An introduction to stochastic processes." Cambridge Univ. Press (1956).
5. Parzen, E.: An approach to time series analysis. Ann. Math. Stat. **32,** 951–988 (1961).
6. Kaula, W. M.: Statistical and harmonic analysis of gravity. *J. Geophys. Res.* **65,** 2401–2421 (1959).
7. Kozai, Y.: The potential of the earth derived from satellite motions. Proc. IUTAM Sym. Dynamics of Satellites, Paris (1962).
8. Moe, K.: The errors in orbital predictions for artificial earth satellites. *Pres. Amer. Geophys.* Un. 43d. Ann. Meeting, Washington (1962).
9. Kaula, W. M.: Analysis of satellite observations for longitudinal variations of the gravitational field. *Space Research II, Proc. 2d Int. Space Sci. Sym.* 360–372 (1961).
10. ———: Satellite orbit analysis for geodetic purposes. Proc. Conf. Gen. & Prac. Problems in Th. Astron., Moscow (1961).
11. Jacchia, L. G.: A variable atmospheric density model from satellite observations. *J. Geophys. Res.* **65,** 2775–2782 (1960).

TESSERAL HARMONICS OF THE GRAVITATIONAL FIELD AND GEODETIC DATUM SHIFTS DERIVED FROM CAMERA OBSERVATIONS OF SATELLITES

WILLIAM M. KAULA

Goddard Space Flight Center

Baker–Nunn camera observations of satellites 1959α1 over 315 days, 1959 η over 105 days, and 1960ι2 over 294 days were analyzed for 35 spherical harmonic coefficients of the earth's gravitational field and for position shifts of six geodetic datums. Of the three satellites, only 1960ι2 appeared to have a sufficiently good observation distribution and small enough drag effects to yield significant results.

The datum shifts obtained have standard deviations averaging ± 25 m in each coordinate. The gravitational harmonic coefficients obtained appear to be appreciably different from zero for indices (n,m) at (2,2), (3,1), (4,1) and (4,3). In particular, geophysically significant magnitudes were obtained for J_{22}: 2.51×10^{-6} and J_{31}: 1.79×10^{-6} (normalized).

INTRODUCTION

The Baker–Nunn camera observations described by Veis[1,2] and Haramundanis[3] will be analyzed by the methods described in references 4–7. Solutions were made for all geodetic and gravitational parameters estimated to have effects of more than ± 20 m on satellite orbits. The intent of the analysis was to apply all devices short of allowing for covariance of observations at different times. This intent resulted in programs complicated enough that most of the time spent was consumed by purely computational difficulties. An IBM 7090 computer was used.

OBSERVATIONS

The Baker–Nunn camera system, its accuracy, and operation by the Smithsonian Institution Astrophysical Observatory are described by Henize,[8] Lassovszky,[9] Weston,[10] Veis and Whipple.[11] That the random error of the plate measurements is of the order of $\pm 2''$ has been confirmed in this analysis by accuracy with which a line can be fitted to plotted residuals with respect to an orbit of observations close together in the same pass. Since the significant timing error is virtually constant throughout a pass, no such test of timing errors is possible because of the dominant effect of drag error in the orbit.

The Baker–Nunn camera observations as published by Veis[1,2] and Haramundanis[3] are referred to the 1950 mean positions of the stellar catalog. For this analysis, the epoch of the right ascension and declination was updated to the epoch of the orbital arc fitted to the observations, taking into account precession plus nutational terms of more than $0''.25$ amplitude—i.e., the 18.6 year and semiannual terms. A.1 times are given for the observations and are treated as equivalent to ephemeris time. A small correction was applied in calculating Greenwich Sidereal Times (GST) to allow for the precession and nutation between the epoch of the orbital arc and the instant of observation.

The above mentioned $\pm 2''$ accuracy of fitting of a line to residuals is appreciably smaller than the residuals themselves, which indicates that extra observations within a pass did not add extra weight to the orbit analysis. Hence, to conserve computer time and to avoid overweighting certain passes, observations were omitted which were neither terminal observations of a pass nor ob-

servations interior to a pass at intervals of 2 minutes or more.

The final rejection criterion applied was to omit observations on days of appreciable atmospheric disturbance, as measured by the geomagnetic index A_p. For the 1960ι2 (Echo I rocket body) analysis, observations were omitted on days for which A_p exceeded 50; for 1959α1 (Vanguard II) and 1959η (Vanguard III), when A_p exceeded 70. In some cases additional observations on adjacent days were omitted to prevent an orbital arc from bridging across days of high A_p index.

The principal defect in the observations is, of course, their poor distribution due to the dependence on reflected sunlight; to the limited number of tracking stations—twelve; and, in the case of 1959α1 and 1959η, to the closeness of the satellite perigees.

GEOMETRY

The observation equation used was in terms of the meridian and prime vertical components of the plate measurement, assuming that the satellite was on the camera axis,[4,12] and consists of the first two rows of the matrix equation:

$$\begin{Bmatrix} d\delta \\ d\alpha \cos\delta \\ \frac{dr}{r} \end{Bmatrix} = \left\{\frac{\mathbf{db}}{r}\right\}_{obs} = -\left\{\frac{\mathbf{b}}{r}\right\}_{obs} + \frac{\mathbf{R}_{bx}\{\mathbf{R}_{xq}\,\mathbf{q} + \mathbf{C}_{xe}\,d\mathbf{e} + \mathbf{C}_{xM}\,n\,dt - \mathbf{R}_3(-\theta)\,d\mathbf{u}_0\}}{r}, \quad (1)$$

where δ is the declination, α the right ascension, r the camera-satellite range. In equation 1, the first two rows of $(\mathbf{b}/r)_{obs}$ are zero, if the observed δ, α are used in $\mathbf{R}_{bx}$, the rotation matrix from the inertial coordinate system to a rectangular system with the 3-axis coincided with the camera-satellite line, and the 1-axis in the meridian; $\mathbf{q}$ is the satellite position in orbit-referred coordinates, with the 1-axis toward osculating perigee and the 3-axis normal to the osculating orbit; $\mathbf{R}_{xq}$ is the rotation from orbit-referred to inertial coordinates; $\mathbf{C}_{xe}$ is a 3×6 matrix of partial derivatives of the inertial rectangular coordinates with respect to the osculating Keplerian elements, corrections to which are symbolized by $d\mathbf{e}$; $\mathbf{C}_{xM}$ is the row of $\mathbf{C}_{xe}$ corresponding to the mean anomaly; n is the mean motion; dt is a correction to the time of observation; $\mathbf{R}(-\theta)$ is the geodetic to inertial rotation matrix, with the Greenwich Sidereal Time θ as argument; and $d\mathbf{u}_0$ is a vector of corrections to station position. (Derivations of all these variables are given in equations 46, 47, 52–60 of reference 4, or equations 3.1–3.8, 3.11–3.15 of reference 12.)

The partial derivatives in equation 1,

$$\mathbf{C}_t = \begin{Bmatrix} \frac{\partial\delta}{\partial t} \\ \frac{\partial(\alpha\cos\delta)}{\partial t} \end{Bmatrix} = \begin{Bmatrix} 1, & 0, & 0 \\ 0, & 1, & 0 \end{Bmatrix} \frac{\mathbf{R}_{bx}\mathbf{C}_{xM}}{r}\, n, \quad (2)$$

were not actually used to determine timing corrections; but were used for three other purposes: (1) A correction $r\mathbf{C}_t/c$ was applied for the time of travel of the signal (c is the velocity of light); (2) a lower weight was given to the along-track component than to the across-track component of the observation, by giving each observation a 2×2 covariance matrix,

$$\mathbf{V}_{obs} = \begin{Bmatrix} \sigma_d^2 & 0 \\ 0 & \sigma_d^2 \end{Bmatrix} + \mathbf{C}_t\,\sigma_t^2\,\mathbf{C}_t^T, \quad (3)$$

where σ_d^2 is the variance of the direction measurement, σ_t^2 is the variance of the timing, and the superscript T denotes transpose; and (3) the residuals in along-track and cross-track components were computed by applying to $(\delta, \alpha\cos\delta)$ residuals the rotation:

$$\mathbf{R}_{ti} = \begin{Bmatrix} \frac{C_1}{\sqrt{C_1^2+C_2^2}}, & \frac{C_2}{\sqrt{C_1^2+C_2^2}} \\ \frac{-C_2}{\sqrt{C_1^2+C_2^2}}, & \frac{C_1}{\sqrt{C_1^2+C_2^2}} \end{Bmatrix}, \quad (4)$$

where C_1, C_2 are the two elements of $\mathbf{C}_t$.

Consistent with the assumption, stated in the introduction, of seeking all effects expected to be larger than ±20 m, all stations were assumed to have position error, but those stations connected

TABLE 1.—*Tracking Station Data in Length Units of* 6.378165 *m*

Station	Latitude and Longitude (degrees)	Datum	Starting Coordinates	Preassigned σ	Solution	
					Preliminary	Final
Organ Pass	32.4	Americas	-240778.9	± 3.0	-14.8	- 2.8 ± 0.9
	253.4		-810109.7	± 3.9	- 5.6	- 3.8 ± 1.8
			+533234.2	± 3.1	+ 3.3	- 0.2 ± 1.0
Arequipa	-16.5		+304591.7			
	288.5		-909989.8			
			-281725.5			
Curacao	12.1		+353051.9			
	291.2		-912004.8			
			+208079.7			
Jupiter	27.0		+153068.1			
	279.8		-878214.3			
			+451581.1			
Olifants-Fontein	-26.0	Europe-Africa-Siberia-India	+792726.2	± 3.4	+16.3	+ 5.4 ± 1.3
	28.3		+425915.7	± 2.9	-17.8	- 8.0 ± 1.4
			-435196.6	± 2.9	+ 0.1	+ 2.6 ± 2.5
San Fernando	36.5		+800481.9			
	353.8		- 87033.6			
			+591042.3			
Naini Tal	29.4		+159627.8			
	79.5		+857813.2			
			+487527.7			
Shiraz	29.6		+529444.8			
	52.5		+690490.0			
			+491723.2			
Woomera	-31.1	Australia	-624562.7	±11.3	+ 6.4	-15.0 ± 4.0
	136.8		+586884.9	±14.5	+14.5	+ 4.4 ± 8.3
			-513573.3	±13.2	+ 2.8	+ 8.7 ± 6.5
Tokyo	35.7	Japan-Korea-Manchuria	-618774.5	± 5.2	0.0	- 8.5 ± 2.7
	139.5		+527787.3	± 6.9	+15.2	+ 6.8 ± 3.4
			+579917.5	± 5.2	+ 1.9	+ .9 ± 2.4
Villa Dolores	-31.9	Argentina	+357509.6	±28.4	+27.1	+36.9 ± 3.9
	294.9		-770550.4	±22.8	- 2.2	+ 3.4 ± 5.0
			-526083.5	±26.2	+ 1.2	+ 3.3 ± 6.6
Maui	20.7	Hawaii	-857008.8	±21.7	+ 1.8	+ 1.5 ± 6.1
	203.7		-376954.1	±35.8	+28.5	+14.1 ± 7.9
			+351587.3	±37.1	-53.3	-50.1 ± 4.7

to the same geodetic system were assumed to shift together. Hence the twelve cameras were referred to six datums: four to the Americas (Am) system; four to the Europe–Africa–Siberia–India (EASI) system; and one each to the Australian (Au), Japan–Korea–Manchuria (JKM), Argentine (Ar), and Hawaiian (H) systems. For the Am, EASI, JKM systems, the starting station positions were those obtained in the solution for a world geodetic system of Kaula.[13] For the Au, Ar, and H systems the positions calculated by Veis[14] were taken and shifted by placing tangent to the datum origin an ellipsoid of flattening 1/298.3 and an equatorial radius of $6378165+N_0$ m, where N_0 is the geoid height in the vicinity of the datum origin as given by Kaula.[13] The initial station positions are given in table 1, in length units of 6.378165 m referred to the **U** coordinate system, with axes toward 0°, 0°; 0°, 90° E; and 90° N, respectively.

DYNAMICS

Variables in the observation equation (equation 1) dependent on the dynamics of the satellite orbit are:

$$\mathbf{R}_{xq} = \mathbf{R}_3(-\Omega)\,\mathbf{R}_1(-i)\,\mathbf{R}_3(-\omega) \qquad (5)$$

and

$$\mathbf{q} = \begin{Bmatrix} a(\cos E - e) \\ a\sqrt{1-e^2}\sin E \\ 0 \end{Bmatrix} \qquad (6)$$

where E, a, e, i, ω, Ω are the osculating eccentric anomaly, semi-major axis, eccentricity, inclination, argument of perigee, and longitude of the ascending node, respectively; and

$$d\mathbf{e} = \mathbf{J}\,d\mathbf{e}_0' + \mathbf{C}_{eg}\,d\mathbf{p}_g + \mathbf{C}_{et}\,d\mathbf{p}_t + \mathbf{C}_{ed}\,d\mathbf{p}_d + \mathbf{C}_{ep}\,d\mathbf{p}_p \qquad (7)$$

where $\mathbf{e}_0'$ denotes the elements of an intermediate orbit at epoch; $\mathbf{p}_g$, parameters expressing variations in the earth's gravitational field (such as spherical harmonic coefficients); $\mathbf{p}_t$, arbitrary polynomials of the Keplerian elements; $\mathbf{p}_d$, parameters of an atmospheric model and the interaction therewith of the satellite; and $\mathbf{p}_p$, parameters expressing radiation pressure effects.

The procedure used to compute the osculating elements M, a, e, i, ω, Ω and the partial derivatives matrices $\mathbf{J}$, $\mathbf{C}_{eg}$, $\mathbf{C}_{et}$, $\mathbf{C}_{ed}$, $\mathbf{C}_{ep}$ was as follows:

Preliminary orbits were determined by iterated differential correction fit to the observations based on the parameters: (1) The constants of integration of the orbital theory of Brouwer[15]; (2) the gravitational field parameters kM, and zonal harmonics J_2, J_3, J_4; and (3) arbitrary polynomials in time of the Keplerian elements. The principal purpose of this preliminary orbit determination was to obtain osculating elements at the instant of each observation close enough to the true values that the corrections could be considered linear.

The intermediate orbit elements defining the preliminary orbit were used to generate Fourier series expressing the effects of the several perturbations and the partial derivatives of the osculating elements with respect to the parameters of the perturbations.

For $\mathbf{C}_{eg}$, the effect of spherical harmonics of the earth's gravitational field, the disturbing function developed in Kaula[4] was used:

$$R_{nm} = \frac{a_e^n \mu}{a^{n+1}} \sqrt{\frac{(n-m)!\,(2n+1)\,\kappa_m}{(n+m)!}} \sum_{p=0}^{n} F_{nmp}(i) \sum_{q=-\infty}^{\infty} G_{npq}(e)$$

$$\cdot \left[\begin{Bmatrix} \bar{C}_{nm} \\ -\bar{S}_{nm} \end{Bmatrix}_{(n-m)\text{ odd}}^{(n-m)\text{ even}} \cos\{(n-2p)\omega + (n-2p+q)M + m(\Omega-\theta)\} + \begin{Bmatrix} \bar{S}_{nm} \\ \bar{C}_{nm} \end{Bmatrix}_{(n-m)\text{ odd}}^{(n-m)\text{ even}} \{\sin (n-2p)\omega + (n-2p+q)M + m(\Omega-\theta)\} \right], \qquad (8)$$

where $\kappa_0 = 1$; $\kappa_m = 2$, $m \neq 0$.

TABLE 2.—*Gravitational Coefficient Data (All Numbers Multiplied by a Scaling Factor of 10^{-6})*

Gravitational Coefficient*	Starting Value	Preassigned σ	Solution		Gravitational Coefficient*	Starting Value	Preassigned σ	Solution	
			Preliminary	Final				Preliminary	Final
$\Delta \bar{C}_{00}$ †	.00	±10.00	4.52	1.23 ±3.29	$\bar{C}_{43}$	.00	± .63	1.23	.50 ± .21
$\Delta \bar{C}_{20}$ †	-.054	± .07	.01	- .06 ± .01	$\bar{S}_{43}$	.00	± .63	.10	.16 ± .19
$\bar{C}_{21}$	.00	± 2.00	- .04	Fixed	$\bar{C}_{44}$	.00	± .63	- .48	- .24 ± .27
$\bar{S}_{21}$	.00	± 2.00	- .07	Fixed	$\bar{S}_{44}$	.00	± .63	1.12	.55 ± .29
$\bar{C}_{22}$	.00	± 2.00	2.96	1.84 ± .19	$\bar{C}_{50}$	.019	± .02	- .30	.03 ± .01
$\bar{S}_{22}$	.00	± 2.00	-1.71	-1.71 ± .28	$\bar{C}_{51}$	.00	± .39	.35	.08 ± .14
$\bar{C}_{30}$	.970	± .02	1.07	.98 ± .01	$\bar{S}_{51}$	.00	± .39	.70	.26 ± .15
$\bar{C}_{31}$	.00	± 1.26	1.89	1.77 ± .21	$\bar{C}_{60}$	-.110	± .10	.00	- .10 ± .02
$\bar{S}_{31}$	.00	± 1.26	.28	- .11 ± .20	$\bar{C}_{61}$	.00	± .28	.06	.02 ± .08
$\bar{C}_{32}$	.00	± 1.26	- .37	.34 ± .26	$\bar{S}_{61}$	.00	± .28	- .05	- .18 ± .07
$\bar{S}_{32}$	.00	± 1.26	.30	.08 ± .35	$\bar{C}_{62}$	.00	± .28	.21	.00 ± .06
$\bar{C}_{33}$	.00	± 1.26	- .32	- .31 ± .46	$\bar{S}_{62}$	.00	± .28	.07	.06 ± .07
$\bar{S}_{33}$	.00	± 1.26	.61	.74 ± .46	$\bar{C}_{63}$	.00	± .28	1.11	.13 ± .08
$\bar{C}_{40}$	.613	± .12	.79	.55 ± .10	$\bar{S}_{63}$	.00	± .28	.31	.21 ± .16
$\bar{C}_{41}$	.00	± .63	- .10	- .21 ± .16	$\bar{C}_{64}$	.00	± .28	.00	.13 ± .11
$\bar{S}_{41}$	.00	± .63	.60	.46 ± .15	$\bar{S}_{64}$	.00	± .28	- .28	- .24 ± .10
$\bar{C}_{42}$	.00	± .63	.66	- .03 ± .19	$\bar{C}_{70}$	.121	± .02	- .13	.10 ± .01
$\bar{S}_{42}$	.00	± .63	.25	.32 ± .19					

* $\bar{C}_{nm}$, $\bar{S}_{nm}$ are coefficients of the spherical harmonic terms $kM/r\,(a_e/r)^n H_{nm}$ such that $\int H_{nm}^2\, d\sigma = 4\pi$ for integration over the sphere (Reference 17, Equations 16-18).

† $\Delta\bar{C}_{00}$, $\Delta\bar{C}_{20}$ are corrections to $0.3986032 \times 10^{21}\,(1.0 - 0.00108236\,P_2)$ cgs.

This disturbing function was used in the Lagrangian equations of motion (reference 16, p. 289) and integrated under the assumption that a, e, i remained constant and M, ω, Ω changed secularly. The program automatically determined for each spherical harmonic all terms above a specified minimum, in absolute magnitude, and stored the results as subscripted numerical arrays to be multiplied by the sines and cosines evaluated at the instant of each observation. An example of one of the 210 such partial derivatives formed for satellite 1960ι2 is

$$\frac{\partial e}{\partial \bar{C}_{31}} = 1.850\cos(\omega+\Omega-\theta) - 0.001\cos(\omega+M+\Omega-\theta) + 5.058\cos(-\omega+\Omega-\theta)$$

$$+ 0.002\cos(-\omega-M+\Omega-\theta) - 0.609\cos(-\omega-2M+\Omega-\theta) . \quad (9)$$

By using a rejection criterion of $0.1n^{1.2}$ and applying it to partial derivatives of the elements $M+\omega+\Omega\cos i$, $e^2(\omega+\Omega\cos i)$, $\Omega\sin i$, e, i, and a between 1 and 6 significant periodicities were found for each term.

The harmonics listed in table 2 were selected on the basis that they has a rms anticipated effect on the satellite orbit of $\pm 20m$ or more, using the degree variances given in reference 17.

As expected, the partial derivatives indicated poor separation of even degree harmonics of the same order m. However, the effect of the different frequency odd—especially third—degree harmonics was unexpectedly distinct. The even degree harmonics caused principally along-track perturbations of frequency $m(\dot{\Omega}-\dot{\theta})$, while the odd degree harmonics perturbed mainly the eccentricity (or perigee height) for a nearly circular orbit.

For tesseral harmonic coefficients, initial values of zero were assumed; for zonal harmonic coefficients, the values of Kozai[18] were used. For the gravitational effects of the sun and moon, the similar disturbing function in reference 19 was used. All secular terms were retained, plus periodic terms of more than 2×10^{-5} amplitude, of which 2 to 9 were found for each orbit. For the radiation pressure effect of the sun, the disturbing function in reference 19 was used. Because of the irregular effect of the earth's shadow, the perturbations were not integrated analytically, and a numerical harmonic analysis was applied instead. A harmonic analysis interval of 15 days (or minimum period of 30 days) was found sufficient to reflect all variations of more than 2×10^{-5} amplitude. Partial derivatives were formed only for one parameter: the mean (reflectivity X cross-sectional area).

For drag, the effect of an empirical atmospheric model was applied with density in the form[20]:

$$\rho = \rho_0 \left(\frac{S}{100}\right)^m \exp\left\{\frac{h-h_0}{H} + ce^{-dh}\right\}\left\{1 + b\left(e^{ah}-k\right)\cos^n\frac{\psi}{2}\right\} . \quad (10)$$

In equation 10, S is the solar flux of 10.7 (or 20) cm wavelength, h is the height above the earth's surface, and ψ is the angle from the center of the diurnal bulge and is determined by

$$\cos\psi = \frac{\{1,\ 0,\ 0\}\,R_3(\lambda^*)\,R_1(\epsilon)\,R_3(\chi)\,R_{xq}\,q}{r} , \quad (11)$$

where λ^* is the sun's longitude, ϵ the inclination of the ecliptic, and χ is the lag of the atmospheric bulge behind the sun.

The atmosphere was assumed to rotate with the solid earth, and to have the corresponding oblateness for a fluid. The customary assumption of the drag force being proportionate to the square of the velocity was made. The force components (radial, transverse, and normal to the satellite and its orbital plane) were used in the Gaussian equations of motion (reference 16, p. 301) and in numerical Fourier series developed for the effects on the Keplerian elements. In generating these series, second-order effects on the angular elements dependent on the secular motions due to the oblateness were included. With an analysis

TABLE 3.—*Satellite Orbit Specifications*

Orbital Data*	Satellite		
	1960 ι2	1959 α1	1959 η
Epoch	1960 Sept. 22.0	1959 Feb. 28.5	1959 Sept. 28.5
Semi-major axis, a (units of 6378165.m)	1.250057	1.304585	1.334500
Eccentricity, e	.011459	.1658200	.1900819
Inclination, i	.8243362	.5738098	.5821184
Argument of Perigee, ω	2.263771	3.360617	3.204033
Longitude of Node, Ω	2.281389	2.524415	3.483041
Mean anomaly, M	2.728678	6.004626	3.824077
Perigee motion/day, $\dot{\omega}$	+.051863	+.091813	+.085014
Nodal motion/day, $\dot{\Omega}$	-.054127	-.061077	-.057119
Maximum Area-to-Mass ratio (cm^2/gm)	0.27	0.21	0.27
Minimum Area-to-Mass ratio (cm^2/gm)	0.08	0.21	0.04
Perigee height (km)	1500	560	510

*The six orbital elements at epoch are the constants of integration as defined by Brouwer's theory (Reference 16).

interval of 3 days, variations as small as 3×10^{-6} amplitude were obtained in M.

For satellites 1959α1 and 1959η the values of the parameters in equation 9 as determined by Jacchia[20] were used. For satellite 1960ι2, c, a, and k were set equal to zero, and ρ_0, m, H, b, n, and χ were determined so as to fit the atmospheric models of Harris and Priester.[21] For 1959α1 and 1959η, the Jacchia model absorbed most of the long period drag variations, but did not fit variations characterized by periods of less than 10 days. For 1960ι2, the Harris and Priester model did not reduce residuals significantly, and had a negligible effect on the values determined for the geodetic parameters; hence the model was omitted.

In computing the effects of arbitrary polynomials or the partial derivatives with respect thereto, the second-order effects of the acceleration based on the assumption of constant perigee height (equations 5–14 of reference 22; equation 2.100 of reference 12) were applied.

In the partial derivatives **J** with respect to the intermediate orbital elements at epoch (equation 7), the effects of secular motions due to oblateness were included (reference 4, Equation 49). To assure that the $\pm20m$ specification was met, the extension of Brouwer's theory to periodic terms of order J_2^2 by Kozai[23] was examined but was found not to be needed.

In the final orbit analysis, the various perturbations were added to the osculating elements as

determined from the preliminary orbit at each observation. To keep these preliminary positions close to the actual positions, the longer period drag and radiation pressure effects were subtracted from the arbitrary polynomials of the preliminary orbit. Provision was also made to eliminate any specified polynomials in the final orbit determination.

Orbital elements for the initial epoch, and other specifications, of each satellite are given in table 3.

DATA ANALYSIS

As discussed in references 5, 6, and 7, difficulties are created by: (1) The nonuniform distribution of observations; (2) the similarity of effects on the observations of different gravitational coefficients and station position errors; (3) the inadequacy of the atmospheric model; and (4) the prohibitive amount of computing time which would be required by a solution taking account of serial correlation between different times. Five methods were suggested to overcome these difficulties:

1. Inclusion of all possible significant perturbations by either physical or empirical models.
2. Preassigned variance and covariance **V** for the starting values of parameters for which corrections z are being determined so that the solution becomes[4]:

$$z = \left(\mathbf{M}^T \mathbf{W}^{-1} \mathbf{M} + \mathbf{V}^{-1}\right)^{-1} \mathbf{M}^T \mathbf{W}^{-1} \mathbf{f}, \qquad (12)$$

where **W** is the covariance matrix of the observations, **M** is the matrix of partial derivatives in the observation equations, and **f** is the vector of residuals.

3. Assignment of higher weight to the across-track than to the along-track component of an observation, as described by equation 3.
4. Observations weighted inversely as their density with respect to phase angles critical for determination of the geodetic parameters, such as node-GST.
5. Use of arbitrary polynomials.

To this list we could add lower weighting—or omission—of observations on days of considerable atmospheric irregularity, according to some index such as the aforementioned A_p.

The application of the first of these methods is described in the section on dynamics. It was found that inclusion or omission of effects which were secular or of periods more than a few days had very little influence on the values determined for the station shifts or tesseral harmonics. The most troublesome inadequacy was the inability of the empirical atmospheric models to explain orbital variations in the 1.0- to 0.1-cycle-per-day part of the spectrum. The principal improvement possible would be to utilize the correlation of corpuscularly caused density variations with the A_p index.[24]

The second device, specifying variance and covariance for the starting values of the parameters, was necessary to avoid absurdly distorted results due to the ill-conditioning caused by nonuniform distribution of observations coupled with the inadequate accounting for drag effects. For the stations on the Am, EASI, and JKM geodetic systems, the 9×9 covariance matrix generated in the solution of reference 13 was used. For the three isolated datums, the assigned covariance matrices were based on assumed error ellipsoids with $\pm35m$ vertical semi-axes in all three cases, and horizontal semi-axes of $\pm100m$ for Au; $\pm200m$ for Ar; and $\pm250m$ for H. The smaller uncertainty for the Australian system is based on the improvement of its position obtained by adjusting deflections-of-the-vertical.[14] For the zonal spherical harmonic coefficients of the gravitational field, the preassigned variances were based on four times the uncertainties given by Kozai.[18] For the tesseral harmonics $n,m=2,1$ and 2,2, the preassigned variance of $(2.0\times10^{-6})^2$ was based on the order-of-magnitude of earlier determinations of J_{22} by Kozai,[25] Kaula,[5] and Newton.[26] For the tesseral harmonic coefficients of the third and higher degrees the preassigned σ's in table 2 were computed from the degree variances $\sigma_n^2\{\Delta g\}$ in reference 17:

$$\sigma^2\left\{\bar{C}_{nm} \text{ or } \bar{S}_{nm}\right\} = \frac{\sigma_n^2\{\Delta g\}}{(n-1)^2 g^2 (2n+1)}. \qquad (13)$$

Probably, the principal defect of the preassigned variance technique is a tendency to prejudge the results, i.e., if two or more parameters have similar effects on the orbit, the solution will tend to be an overestimate for those corrections whose absolute magnitudes are smaller than ex-

pected, and an underestimate for those corrections larger than expected. In cases where a series of observations of a satellite were referred to several orbital arcs, the datum and gravitational coefficient variances and covariances were multiplied by the number of orbital arcs, so that the preliminary estimates do not have excessive influence on the final mean value.

The influence of the parameters' preassigned variances and covariances on the result does not depend on their absolute magnitude but rather on their magnitude relative to the variances and covariances specified for the observations. This effect is a consequence of the quadratic sum minimization implied by equation 12:

$$S = x^T W^{-1} x + z^T V^{-1} z \,, \qquad (14)$$

where x is the vector of corrections to observations. If the variance assigned to the observations is that based on estimates of purely observational accuracy: direction $\pm 2''$ (reference 9); and timing $\pm 0.002^s$ (reference 10), then grossly distorted values will be obtained for the geodetic parameters. This distortion is the result of neglecting drag effects. If the covariance, from this drag, between observations a few days apart is not taken into account, the next best solution would be to increase the variance of each observation. The amount of the increase must be found by trial and error. Thus, for example, for satellite 1960ι2, initially variances were used of $(20'')^2 = (10^{-4})^2$ for direction and $(0.056^s)^2$ for timing over 23-day arcs. Solving a single orbit for all the parameters resulted in rms residuals of $\pm 4.5'' = \pm 2.2 \times 10^{-5}$ across track and $\pm 9.7'' = \pm 4.9 \times 10^{-5}$ along track. Since the mean range of the satellite is 2100 km/sec and its motion at right angles to the line of sight (equivalent to the $\mathbf{C}_t$ of equation 3) averaged about 5.5 km/sec, and since all of the $\pm 2.2 \times 10^{-5}$ across-track residual is ascribed to directional error, the timing error is

$$\sigma(t) \approx \frac{\bar{r}}{n} \sqrt{\sigma^2_{along} - \sigma^2_{across}} = \frac{2100}{5.5} \sqrt{4.9^2 - 2.2^2} \times 10^{-5} = \pm 0.017^s \,.$$

A direction variance of $(2.2 \times 10^{-5})^2$ and a timing variance of $(0.017^s)^2$ were applied in analyzing all thirteen 23-day orbital arcs of 1960ι2. The resulting corrections to the geodetic parameters are appreciably larger than expected and are listed as "preliminary results" in tables 1 and 2. Readjusting orbital elements and arbitrary polynomials while holding geodetic parameters fixed obtained rms residuals of $\pm 13.2'' = \pm 6.45 \times 10^{-5}$ across-track and $\pm 28.5'' = \pm 1.38 \times 10^{-4}$ along-track. Hence the small residuals for the single 23-arc had been obtained at the expense of distorted geodetic parameters. The observation variances were changed to $(6.45 \times 10^{-5})^2$ direction and

$$[2100(1.38^2 - 0.645^2)^{1/2} \times 10^{-4}/5.5]^2 = (0.046^s)^2$$

timing for a repeated analysis to obtain the results listed as "final" in tables 1 and 2.

Table 4 shows the distribution of observations with respect to some angles which appear as arguments of the principal terms in the tesseral harmonics effects. In table 3 the observations are sorted into 24 sets. The n_i observations in the ith set all have the pertinent angle between $\pi(i-1)/12$ and $\pi i/12$. The distribution is far from ideal, but applying weighting factors $\sum_i n_i/24n_j$ appeared to have little effect on the results. For the final analysis of the 1960ι2 orbit, weighting with respect to the angle node-GST was applied.

Application of the final device, arbitrary polynomials, was limited to the mean anomaly and to the fourth power in time to avoid the ill-conditioning which occurs in determining the coefficients of higher degree power series. Adding a specification that the orbital residuals should not average more than, say, ten times the $\pm 20m$ minimum amplitude effect sought sets a limit on the length of the orbital arcs which can be treated. This limit appeared to be about 23 days for 1960ι2. The 21-day arcs used for satellites 1960α1 and 1960η resulted in residuals of $\pm 3.6 \times 10^{-4}$ across-track and $\pm 4.7 \times 10^{-4}$ along-track, which are appreciably in excess of the limit. However, the 21-day arcs averaged only 35 observations each, and shorter arcs would have had so few observations that an excessive amount of the

TABLE 4.—*N_i, Counts of Observation Distribution with Respect to Orbital Angles*

Satellite Angle (degrees)	Satellite Observations				
	1960 ι2 (Ω- GST)	1960 ι2 (ω+Ω- GST)	1960 ι2 (- ω + Ω- GST)	1959 α1 (Ω- GST)	1959 η (Ω- GST)
0 - 15	53	40	93	7	2
15 - 30	66	69	48	6	1
30 - 45	67	104	88	6	3
45 - 60	71	97	61	19	17
60 - 75	64	65	65	24	11
75 - 90	89	57	77	19	7
90 - 105	54	56	69	30	12
105 - 120	76	42	74	27	13
120 - 135	41	31	59	46	9
135 - 150	57	47	65	55	14
150 - 165	78	57	89	50	8
165 - 180	66	86	68	27	3
180 - 195	91	56	59	52	19
195 - 210	76	61	47	45	12
210 - 225	92	84	41	31	27
225 - 240	47	66	58	52	23
240 - 255	79	92	53	46	16
255 - 270	79	70	93	58	8
270 - 285	67	55	60	29	11
285 - 300	52	58	64	17	13
300 - 315	50	92	65	30	14
315 - 330	55	92	51	20	4
330 - 345	43	50	69	17	3
345 - 360	49	35	46	14	4
Total	1562	1562	1562	730	254

effects of the geodetic parameters would have been absorbed by the elements of the reference orbit.

To combine the results of several orbital arcs, the more rigorous method would be to use the covariance matrix **U** of the corrections to parameters produced by each of the least squares determinations to obtain a generalized weighted mean:

$$\text{Mean } \bar{\mathbf{z}} = \left(\Sigma_i \mathbf{U}_i^{-1}\right)^{-1} \Sigma_i \mathbf{U}_i^{-1} \mathbf{z}_i \ . \qquad (15)$$

However, the mean solutions obtained in this manner were so distorted that they merely served as a forceful reminder of the falseness of the basic assumption of randomness of observation errors. Hence the solutions given in tables 1 and 2 were calculated simply by ascribing to each arc a weight proportionate to the total number of observations, in the case of the gravitational coefficients; and proportionate to the number of observations from the datum, in the case of datum shifts. The uncertainties given in the tables are standard deviations computed in the customary manner from the scatter of the results for the different orbital arcs about the mean.

RESULTS AND CONCLUSIONS

In view of the excessive number of determinations of tesseral harmonics already published, it was resolved not to publish any results which did not satisfy the test that small values were obtained for the harmonics $\bar{C}_{21}$, $\bar{S}_{21}$, known to be vanishingly small from latitude variation observations. The analyses of the 1959α1 and 1959η orbits failed this test. The preliminary results of the 1960ι2 analysis passed the test very satis-

TABLE 5.—*Comparison of Results from Different Orbital Arcs of* 1960ι2 *for Some Datum Shifts in Units of* 6.378165*m*

Orbit Number	Argentine Datum				Hawaiian Datum			
	Number of Observations	Δu_1	Δu_2	Δu_3	Number of Observations	Δu_1	Δu_2	Δu_3
1	18	+21.0	+12.8	+14.0	2	-30.6	+27.6	-11.6
2	12	+41.6	-20.7	-34.0	14	+22.2	- 3.0	-43.3
3	4	+89.8	-20.2	+10.3	4	+12.6	+17.1	- 7.9
4	6	+46.4	-19.1	-21.5	6	- 9.9	+24.2	-25.8
5	28	+43.7	+23.4	-20.9	3	-10.0	+11.7	-62.0
6	16	+47.9	+ 0.1	-14.1	8	+23.6	+24.9	-47.4
7	7	+39.7	-14.8	- 0.4	19	+10.6	+13.4	-65.5
8	11	+ 7.7	+ 9.3	+18.1	7	+ 0.5	+48.5	-54.0
9	18	+35.4	+24.6	- 5.8	18	-24.4	-16.5	-64.4
10	24	+45.0	-13.6	+23.5	10	-33.6	+78.0	-52.7
11	11	+57.7	-12.6	+13.7	20	+13.2	- 1.6	-58.5
12	13	+17.0	+ 4.2	+18.4	16	+ 6.5	+17.4	-31.9
13	22	+26.0	+ 8.0	+28.9	0			

factorily, as shown in table 2, and in the final analysis $\bar{C}_{21}$, $\bar{S}_{21}$ were held fixed as 0.

Several results show a convincing consistency from orbit to orbit. Of the 18 datum coordinates, 8 have shifts more than twice their standard deviations while, of the 26 tesseral harmonic coefficients, 7 have magnitudes more than twice their standard deviations. Some of the more marked consistencies, such as the $\bar{C}_{31}$ coefficient, the southward shift of the Hawaiian station, and the eastward shift of the Argentine station are listed in tables 5 and 6. The rms magnitudes of the coefficients for the 3rd, 4th, and 6th degrees are 0.50 to 0.70 of what was expected from the autocovariance analysis—indicating perhaps that the variances of the observations should be decreased.

Except perhaps for the equatorial ellipticity,

TABLE 6.—*Comparison of Results from Different Orbital Arcs of* 1960ι2 *for Some Gravitational Coefficients (All Values Multiplied by a Scaling Factor of* 10^{-6})

Orbit Number	Total Number of Observations	Gravitational Coefficients									
		$\bar{C}_{22}$	$\bar{S}_{22}$	$\bar{C}_{31}$	$\bar{S}_{31}$	$\bar{C}_{41}$	$\bar{S}_{41}$	$\bar{C}_{42}$	$\bar{S}_{42}$	$\bar{C}_{43}$	$\bar{S}_{43}$
1	151	2.50	-0.99	1.86	0.60	-0.50	0.33	1.01	0.55	0.09	0.44
2	161	1.98	-2.64	.61	- .90	.42	1.34	-.35	- .56	.58	- .22
3	73	1.54	-2.90	.75	.28	- .69	.19	-.56	- .26	1.25	.60
4	89	3.54	-1.87	.51	.84	- .33	.86	.98	- .05	- .38	- .14
5	110	1.13	- .68	1.88	- .17	.21	.20	.032	.61	-1.08	-1.76
6	132	2.23	-2.20	1.59	.48	- .90	.07	.14	.22	1.01	.43
7	190	1.82	-1.72	1.65	- .46	.02	.70	.04	.44	1.14	.62
8	109	1.59	-1.07	1.98	- .79	- .53	-.67	-.76	.88	- .07	- .02
9	100	2.33	-1.45	3.73	-1.36	.14	.22	.97	.86	.54	.63
10	118	0.65	-3.94	2.39	.10	-1.31	.72	-.78	-1.06	.029	.21
11	173	1.16	- .80	1.77	.342	- .05	-.04	-.76	1.03	1.46	.83
12	97	2.60	- .27	2.17	.65	.79	1.20	-.065	.97	.82	- .10
13	64	0.90	-2.23	2.72	-1.49	- .38	.67	-.19	.37	.16	- .18

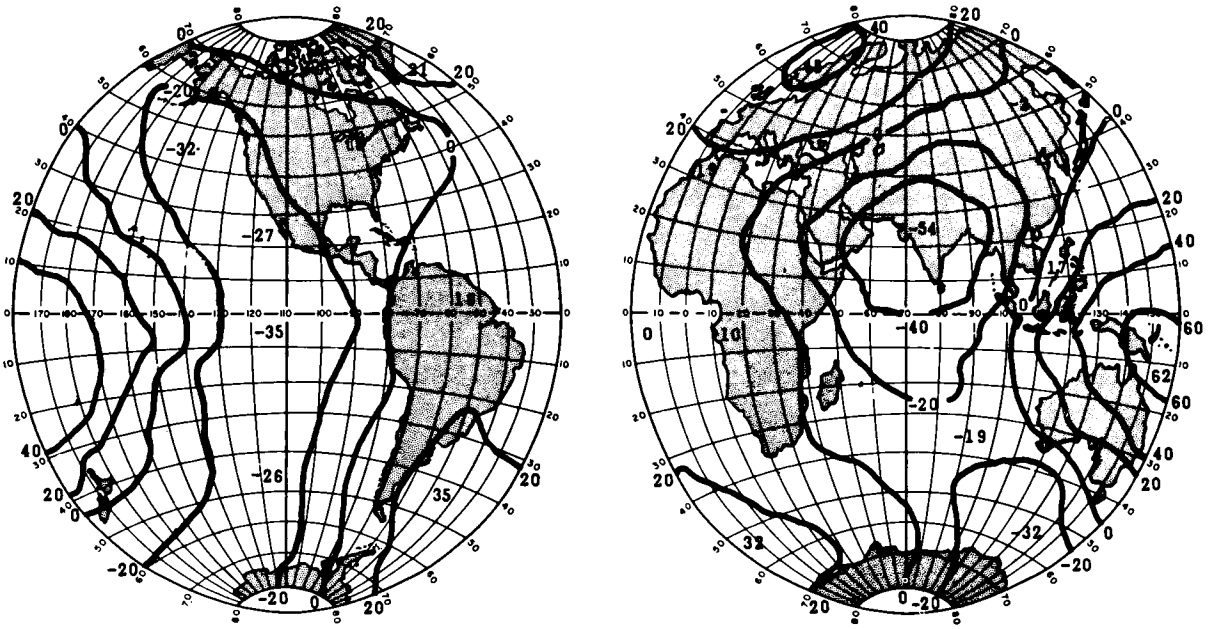


FIGURE 1.—Geoid heights (in meters) referred to an ellipsoid of flattening 1/298.24.

$n,m=2,2$, good agreement with other recent determinations from either satellite or terrestrial data[13,25,27] does not exist. However, for determinations from terrestrial data which are poorly distributed, too much individuality must not be ascribed to the harmonic coefficients; and it is better to make the comparison of spatial, rather than spectral, representations. The geoid in figure 1 corresponds to the final results given in table 2. This satellite geoid agrees quite well in the eastern hemisphere with both the gravimetric geoid of Uotila[27] and the astro-geodetic geoid of Fischer,[28] but in the Western Hemisphere the agreement is poorer.

Undoubtedly, more information of gravitational variations and station positions can be extracted from existing Baker–Nunn camera observations. The question is whether it is worth the computational effort, in view of the current or anticipated satellites which will be more frequently observed by Doppler tracking or optical tracking with artificial illumination. Most of the difficulties encountered in the present analysis could be avoided with a satellite of, say, 1000 km perigee height observed an average of 20 or more times a day. However, a more rigorous statistical analysis of a short arc of Baker–Nunn camera observations by the methods described in reference 6 will be undertaken, as well as the analysis of Doppler observations by the methods described in this paper.

ACKNOWLEDGMENTS

The author gratefully acknowledges Drs. Fred L. Whipple, George Veis, Yoshihide Kozai, and Mr. Imre G. Izsak, all of the Smithsonian Institution Astrophysical Observatory, for advice and provision of results prior to publication. This work was undertaken at the Smithsonian Institution Astrophysical Observatory, but the bulk of it was accomplished at the Goddard Space Flight Center, with the assistance of Mr. Ed Monasterski and Miss Susan Werner.

REFERENCES

1. VEIS, G., "Catalog of Precisely Reduced Observations, No. P–1," Smithsonian Inst., Astrophys. Observ., Res. in Space Sci., Spec. Rept. No. 82, Nov. 30, 1961.
2. ——— "Catalog of Precisely Reduced Observations, No. P–2," Smithsonian Inst., Astrophys. Observ., Res. in Space Sci., Spec. Rept. No. 85, Feb. 12, 1961.

3. HARAMUNDANIS, K., "Catalog of Precisely Reduced Observations, No. P–4," Smithsonian Inst., Astrophys. Observ., Res in Space Sci., Spec. Rept. No. 95, June 18, 1962.
4. KAULA, W. M., "Analysis of Gravitational and Geometric Aspects of Geodetic Utilization of Satellites," NASA Technical Note D–572, March 1961; also *Geophys. J.* **5**(2): 104–133, July 1961.
5. ——— "Analysis of Satellite Observations for Longitudinal Variations of the Gravitational Field," in: *Space Research II: Proc. 2nd Internat. Space Sci. Sympos., Florence, April 1961*, ed. by H. C. van de Hulst, C. de Jager, and A. F. Moore, Amsterdam: North-Holland Publ. Co., 1961, pp. 360–372.
6. ——— Satellite Orbit Analysis for Geodetic Purposes (in Russian), *Bull. Inst. Theor. Astron: Proc. Conf. on General and Practical Problems of Theoretical Astronomy, Moscow* (on press).
7. ——— "Satellite Orbit Analyses for Geodetic Purposes," in: *Proc. Sympos. on the Dynamics of Satellites, Paris, May 1962*, Berlin: Springer-Verlag (in press).
8. HENIZE, K. G., "Tracking Artificial Satellites and Space Vehicles," in: *Advances in Space Science*, ed. by F. I. Ordway III, New York: Academic Press, 1960, vol. 2, pp. 117–142.
9. Lassovsky, K., "On the Accuracy of Measurements Made upon Films Photographed by Baker–Nunn Satellite Tracking Cameras," Smithsonian Inst., Astrophys. Observ., Res. in Space Sci., Spec. Rept. No. 74, Sept. 18, 1961.
10. Weston, E., "Preliminary Time Reduction for the Determination of Precise Satellite Positions," in: *Catalogue of Precise Satellite Positions*, Smithsonian Inst., Astrophys. Observ., Res. in Space Sci., Spec. Rept. No. 41, May 24, 1960, pp. 11–13.
11. VEIS, G., and WHIPPLE, F. L., "Experience in Precision Optical Tracking of Satellites for Geodesy," in: *Space Research II: Proc. 2d Internat. Space Sci. Sympos., Florence, April 1961*, ed. by H. C. van de Hulst, C. de Jager, and A. F. Moore, Amsterdam: North-Holland Publ. Co., 1961, pp. 17–33.
12. Kaula, W. M., "Celestial Geodesy," in: *Advances in Geophysics*, ed. by H. E. Landsberg and J. Van Mieghem, New York: Academic Press, 1962, vol. 9, pp. 191–293.
13. ——— "A Geoid and World Geodetic System Based on a Combination of Gravimetric, Astrogeodetic, and Satellite Data," *J. Geophys. Res.* **66**(6): 1799–1811, June 1961.
14. VEIS, G., "The Positions of the Baker-Nunn Camera Stations," Smithsonian Inst., Astrophys. Observ., Res. in Space Sci., Spec. Rept. No. 59, Mar. 3, 1961.
15. BROUWER, D., "Solution of the Problem of Artificial Satellite Theory Without Drag," *Astronom. J.* **64**(1274): 378–397, November 1959.
16. ———, and CLEMENCE, G. M., "Methods of Celestial Mechanics," New York: Academic Press, 1961.
17. KAULA, W. M., "Statistical and Harmonic Analysis of Gravity," *J. Geophys. Res.* **64**(12): 2401–2421, Dec. 1959.
18. KOZAI, Y., "The Potential of the Earth Derived from Satellite Motions," in: *Proc. Sympos. on the Dynamics of Satellites, Paris, May 1962*, Berlin: Springer-Verlag (in press).
19. KAULA, W. M., "Development of the Lunar and Solar Disturbing Functions for a Close Satellite," NASA Technical Note D–1126, January 1962; also *Astronom. J.* **67**(5): 300–303, June 1962.
20. JACCHIA, L. G., "A Variable Atmospheric-Density Model from Satellite Accelerations," *J. Geophys. Res.* **65**(9): 2775–2782, Sept. 1960.
21. HARRIS, I. and PRIESTER, W., "Time-Dependent Structure of the Upper Atmosphere," NASA Technical Note D–1444, Aug. 1962.
22. O'KEEFE, J. A., ECKELS, A., and SQUIRES, R. K., "The Gravitational Field of the Earth," *Astronom. J.* **64**(7): 245–253, Sept. 1959.
23. KOZAI, Y., "Second-Order Solution of Artificial Satellite Theory Without Air Drag," *Astronom. J.* **67**(7): 446–461, Sept. 1962.
24. JACCHIA, L. G., "Electromagnetic and Corpuscular Heating of the Upper Atmosphere," in: *Space Research III, Proc. 3d Internat. Space Sci. Sympos., Washington, May 1962*, Amerserdam: North-Holland Publ. Co. (to be published).
25. KOZAI, Y., "Tesseral Harmonics of the Gravitational Potential of the Earth as Derived From Satellite Motions," *Astronom. J.* **66**(7): 355–358, Sept. 1961.
26. NEWTON, R. R., "Ellipticity of the Equator Deduced From the Motion of Transit 4A," *J. Geophys. Res.* **67**(1): 415–416, January 1962.
27. UOTILA, U. A., "Corrections to Gravity Formula From Direct Observations and Anomalies Expressed in Lower Degree Spherical Harmonics," Ohio State Univ., Inst. of Geodesy, Photogrammetry and Cartography Rept. No. 23, 1962.
28. FISCHER, I., "An Astrogeodetic World Datum From Geoidal Heights Based on the Flattening $f=1,298.3$," *J. Geophys. Res.* **65**(7): 2067–2076, July 1960.

ORBIT PREDICTION ACCURACY THEORY

H. E. MONTGOMERY, R. C. BLANCHARD, AND F. WHITLOCK

Goddard Space Flight Center

This document presents the development of equations which may be used to estimate the attainable accuracy of an orbit prediction. Given a schedule of observations with the corresponding standard deviations of the measurements and an orbit defined by the orbital elements a, e, i, ω, t_p, Ω the equations of this report may be used to estimate the accuracy to which the orbit may be determined. Specifically, these equations may be used to estimate the variance–covariance matrix of the orbital elements for a specific tracking schedule from which the standard deviations of the orbital elements may be calculated. In addition, the method of calculating the confidence ellipsoid of position and velocity is shown. These are diagonalized matrices whose diagonals give the principal values of the variances in both position and velocity, respectively. The eigenvectors are calculated to give the direction of the ellipsoids with respect to an orbital plane coordinate system. The analysis is done for a variety of observations which include range, range rate, azimuth, elevation, and the minitrack direction cosines. The weighted least squares criterion is used to establish the relationship between the corrections in the orbital elements and the observational errors, that is, the residuals. The required partial derivatives are calculated in closed form from two-body elliptical orbit theory.

INTRODUCTION

This report presents equations for determining orbit prediction accuracy. Mr. R. Lancaster of the Goddard Space Flight Center provided constructure criticism of the theory and its documentation. The original version of this analysis was carried out by Messrs. R. Leach,[2] D. Kraft† and R. Salinger[1] of the Martin Co. (under contract to the Goddard Space Flight Center, contract number NAS5–1611; task order 11).

The duration of the contract did not allow sufficient time for proper documentation and checkout of the equations. Changes and additions have been made to the original analysis to complete the development. Hopefully, this report will serve as a basis for further development of the prediction accuracy program which is now in the "debugging" stage.

NOTATION

$A_1(\alpha)$, $A_2(\alpha)$, $A_3(\alpha)$	Matrices which denote positive rotations through the angle α about axes No. 1, 2 and 3; respectively.
A	Azimuth of the vehicle as observed from a tracking station, defined by equation 27.
a	Semimajor axis.
B	The matrix which transforms a vector in the (X, Y, Z) coordinate system into the (x_s+R, y_s, z_s) coordinate system (see app. A–3).
b_{ij}	ith row and jth column element of B.
C	The column vector of orbital parameters a, e, t_p, ω, Ω, and i, respectively.
cx	$\equiv \cos x$, shorthand notation used in appendix A.
D	$\equiv[x_s\dot{x}_s+y_s\dot{y}_s+z_s\dot{z}_s]/\rho^3$, shorthand notation used in appendix B–2.

† Currently a Goddard employee.

E^*	Eccentric anomaly.
$E(x)$	Expected value of the variable x.
E	Elevation of the vehicle as observed from the tracking station, defined by equation 26.
e	Eccentricity.
$F(x_s, y_s, z_s)$	$[\partial M_i/\partial x_s, \partial M_i/\partial y_s, \partial M_i/\partial z_s]$, the row matrix of the partials of the ith observation with respect to the coordinates x_s, y_s and z_s. These are given in appendix B.
$F(X, Y, Z)$	$[\partial M_i/\partial X, \partial M_i/\partial Y, \partial M_i/\partial Z]$, the row matrix of the partials of the ith observation with respect to the coordinates X, Y, Z. This is calculated by equation 43.
$F(\dot{x}_s, \dot{y}_s, \dot{z}_s)$	$[\partial M_i/\partial \dot{x}_s, \partial M_i/\partial \dot{y}_s, \partial M_i/\partial \dot{z}_s]$, the row matrix of the partials of the ith observation with respect to $\dot{x}_s$, $\dot{y}_s$ and $\dot{z}_s$. These are given in appendix B–2.
$F(\dot{X}, \dot{Y}, \dot{Z})$	$[\partial M_i/\partial \dot{X}, \partial M_i/\partial \dot{Y}, \partial M_i/\partial \dot{Z}]$, the row matrix of the partials of the ith observation with respect to $\dot{X}$, $\dot{Y}$ and $\dot{Z}$. These are calculated by equation 44.
I	The identity matrix.
i	Inclination.
L	Latitude of the tracking station.
M_i	Computed value of the ith observation such as range ρ, range rate $\dot{\rho}_R$, aximuth A, elevation E, or minitrack direction cosines n, l, or m.
m_i	The ith measurement, such as, ρ, $\dot{\rho}_R$, A, E, n, l, or m.
N	The row matrix of the partials of the observations with respect to the orbital parameters times σ_1^{-1}. This matrix forms a row of the Λ matrix. It is defined by equation 36 or 42.
n	Number of observations.
n, l, m	Minitrack direction cosines, defined by equation 25.
n	Mean motion.
P_{pos}	The matrix of the partials of the X, Y, Z with respect to the orbital parameters a, e, t_p, ω, Ω and i. These are given in appendix C–1.
P_{vel}	The matrix of the partials of $\dot{X}$, $\dot{Y}$, $\dot{Z}$ with respect to the orbital parameters a, e, t_p, ω, Ω and i. These are given in appendix C–2.
Q	The variance-covariance matrix of the orbital parameters, defined by equation 13.
$\bar{R}_i$	The ith eigenvector as used in equation 62.
R	The equatorial radius of the earth (6378.165 km).
$\bar{r}$	The geocentric radius vector to the orbiting body.
r	Magnitude of $\bar{r}$.
S	The matrix which transforms a vector in the (x, y, z) coordinate system into the (x_s+R, y_s, z_s) coordinate system (see app. A–2).
sx	$\equiv \sin x$, shorthand notation used in apprneix A.
t	Time.
t_0	Reference time.
t_p	Time of perigee passage.
$U(t)_{\text{pos}}$	The variance–covariance matrix of the variables X, Y and Z. Defined by equation 58.
$U(t)_{\text{vel}}$	The variance–covariance matrix of the variables $\dot{X}$, $\dot{Y}$, $\dot{Z}$. Defined by equation 60.
X, Y, Z	Cartesian coordinates with the origin at the center of the earth, the X-axis in the direction of the radius vector to the satellite, the Z-axis is in the direction of the angular momentum vector and the Y-axis lies in the orbit plane to form a right-hand set.
$\dot{X}', \dot{Y}', \dot{Z}'$	The velocity of the satellite relative to the (X, Y, Z) coordinate system (defined by equation 32).

$\dot{X}, \dot{Y}, \dot{Z}$	The projections of the inertial velocity onto the X, Y, and Z axes, respectively (defined by equation 33).
x, y, z	Cartesian coordinates with the origin at the center of the earth, the x-axis in the direction of the vernal equinox, the z-axis along the north pole of the earth, and the y-axis forms a right-hand set.
x_s, y_s, z_s	Station centered cartesian coordinates assuming a spherical earth, the x_s-axis is normal to the sphere and is positive outward from the surface of the earth, the positive y_s-axis points toward the north and the positive z_s-axis points toward the west.
$\dot{x}$	Time derivative of the x coordinate.
$\cos\alpha, \cos\beta, \cos\gamma$	The minitrack direction cosines (equation 25).
ΔC	The column vector of the incremental changes in the orbital elements, Δa, Δe, Δt_p, $\Delta\omega$, $\Delta\Omega$, and Δi.
ΔE	An incremental change in the eccentric anomaly, E.
ΔM_i	An incremental change of M_i.
Δm_i	$\equiv m_i - M_i]_0$, the residual of the ith observation.
$\Delta X, \Delta Y, \Delta Z$	Incremental changes in X, Y and Z, respectively.
$\Delta\dot{X}, \Delta\dot{Y}, \Delta\dot{Z}$	Incremental changes in $\dot{X}$, $\dot{Y}$ and $\dot{Z}$, respectively.
$\Delta x_s, \Delta y_s, \Delta z_s$	Incremental changes in x_s, y_s and z_s, respectively.
$\Delta\dot{x}_s, \Delta\dot{y}_s, \Delta\dot{z}_s$	Incremental changes in $\dot{x}_s$, $\dot{y}_s$ and $\dot{z}_s$, respectively.
δ_i	The error in the ith observation.
δ	The column matrix of (δ_i/σ_i) where $i=1, 2, \cdots, n$.
Θ	The matrix which transforms a vector in the (x, y, z) coordinate system into the (X, Y, Z) coordinate system (see appendix A–1).
θ	True anomaly (equation 18).
$\dot{\theta}$	Time derivative of θ (equation 21).
Λ	The matrix of the partials of the observations with respect to the orbital parameters times σ^{-1}. It is defined by equation 35.
λ	The right ascension of the meridian of a tracking station at time t (equation 30).
λ_i	The ith eigenvalue as used in equations 61 and 62.
λ_s	The right ascension of the meridian of the tracking station at time t_0.
μ	The earth's gravitational constant (1.407645×10^{16} ft^3/sec^2).
$\bar{\rho}$	The slant range vector. Its components are given by equation 22.
ρ	The magnitude of $\bar{\rho}$ (equation 24).
$\dot{\bar{\rho}}$	The velocity vector of the satellite relative to the tracking station. Its components are given by equation 23.
$\dot{\rho}_R$	The component of $\bar{\rho}$ along $\bar{\rho}$ or the range-rate of the satellite. It is given by equation 29.
σ_i	The standard deviation of the ith observation.
ϕ	The angle from the line of nodes to the satellite measured in the direction of motion or the true anomaly θ plus the argument of perigee ω. It is given by equation 19.
$\dot{\phi}$	Time derivative of ϕ. It is calculated from equation 21.
Ω	The right ascension of the ascending node.
ω	The argument of perigee.
ω_e	The rotation rate of the earth ($\omega_e=7.29211585\times10^{-5}$ rad/sec).

Matrix Operations

A^T	Transpose of the matrix A.
A^{-1}	Inverse of the matrix A.

Subscripts

i ith observation.
j jth orbital parameter.
$x]_0$ x evaluated at reference condition.

THEORY OF ORBIT IMPROVEMENT

Differential Correction Philosophy

The problem of orbit improvement involves using the equations of motion along with observational data to determine the "best" orbit by some appropriate statistical method. Observations calculated from the determined orbit, defined by the classical orbital elements a, e, t_p, ω, Ω, i (or any equivalent set such as x, y, z, $\dot{x}$, $\dot{y}$, $\dot{z}$), differ from the measured observations due to the following:

(1) Approximations involved in the theory of orbit improvement and the mathematical model,
(2) Inaccuracies in the preliminary orbit elements,
(3) Errors in the observation.

Classically, six measurements are sufficient to determine an orbit. If more than six measurements are available, then the additional information can be used to smooth out a portion of the errors. The "best" orbit is obtained in this document by minimizing the residuals (the differences between the observed and calculated measurements) by the weighted least squares technique. This method involves extracting information about the orbit by "smoothing" the errors in the observations. The process is statistical and utilizes an estimate, σ, (the standard deviation, which is a measure of the validity of an observation) to estimate the uncertainties in the orbital elements. The model used is the classical two-body system.

The equations of motion are written as

$$d^2\bar{r}/dt^2 = -\mu(\bar{r}/r^3) \qquad (1)$$

where $\bar{r}$ is the geocentric radius vector from the center of force to the orbiting body, μ is the gravitational constant (1.407645×10^{16} ft^3/sec^2). The solution of equation 1 for observed data m_i is

$$M_i(C_1, C_2, C_3, C_4, C_5, C_6) = m_i \qquad (2)$$

where C_1, C_2, C_3, C_4, C_5, C_6 are the six constants of integration, or the orbital parameters, a, e, t_p, ω, Ω, i involved in the solution of this equation, and M_i is the calculated observation obtained by solving equation 1. The observed data, m_i, can be of the type range, range-rate, azimuth, elevation, direction cosines, or any other measurable quantity. The constants of integration or orbit parameters, C_1, C_2, C_3, C_4, C_5, C_6 could be solved by taking 6 pieces of data, that is, letting $i=1$ to 6. More data would produce redundant information. Since no measurement is exact, an error δ_i can be associated with the ith piece of information. That is,

$$M_i(C_1, C_2, C_3, C_4, C_5, C_6) = m_i - \delta_i \qquad (3)$$

Equations 3 represent the functional relationship which can be simplified by a linear approximation. Expanding equations 3 in a Taylor series about some initial estimate of the orbital parameters, C_{i0}, results in

$$M_i]_0 + \sum_{j=1}^{6} (\partial M_i/\partial C_j)]_0 \Delta C_j = m_i - \delta_i \qquad (4)$$

Defining the residuals by the difference between the observed and computed measurements as

$$\Delta m_i = m_i - M_i]_0$$

equation 4 becomes

$$\Delta m_i = \sum_{j=1}^{6} (\partial M_i/\partial C_j)]_0 \Delta C_j + \delta_i \qquad (5)$$

These are the basic equations of the differential correction technique. The process is to find the change in the orbital parameters, ΔC_j, given the n measurements, m_i. The computed measurement, M_i and the partials of the computed measurement with respect to the orbital parameters $\partial M_i/\partial C_j]_0$ are obtained analytically from the two-body model, given an initial estimate of the orbit parameters, that is, $C_i]_0$ where $i=1$ to 6. The philosophy of orbit improvement might be explicitly stated here as the utilization of more than 6 measurements to give the "best" answers to equations 5. The only parameter in equation 5

to adjust is the measurement errors, δ_i. Consequently, the process is reduced to a "smoothing" of the instrument errors by making

$$(\partial/\partial\Delta C_j)\sum_{i=1}^{n}(\delta_i/\sigma_i)^2=0 \qquad (6)$$

where n is the number of measurements. Each instrument measurement is weighted by σ_i^{-1}, where σ_i is the standard deviation of the ith measurement. The condition given by equation 6 is called the "weighted least square" orbit improvement criterion.

Weighted Least Square Estimate of the Variance-covariance Matrix of the Orbital Elements

Equation 5, developed in the previous section establishes the relationship between the residuals Δm_i's and the incremental change in the orbital elements, ΔC_j's and the instrument errors, δ_i's. Dividing the equation by σ_i, the standard deviation of the ith measurement results in

$$(\Delta m_i/\sigma_i)-\sigma_i^{-1}\sum_{j=1}^{6}(\partial M_i/\partial C_j)]_0\Delta C_j=\delta_i/\sigma_i$$

For n measurements, the equations can be written as column matrices as

$$\begin{bmatrix}\frac{\Delta m_1}{\sigma_1}\\ \frac{\Delta m_2}{\sigma_2}\\ \cdot\\ \cdot\\ \cdot\\ \frac{\Delta m_n}{\sigma_n}\end{bmatrix}-\begin{bmatrix}\frac{1}{\sigma_1}\frac{\partial M_1}{\partial C_1}\Delta C_1+\frac{1}{\sigma_1}\frac{\partial M_1}{\partial C_2}\Delta C_2+\cdots+\frac{1}{\sigma_1}\frac{\partial M_1}{\partial C_6}\Delta C_6\\ \frac{1}{\sigma_2}\frac{\partial M_2}{\partial C_1}\Delta C_1+\frac{1}{\sigma_2}\frac{\partial M_2}{\partial C_2}\Delta C_2+\cdots+\frac{1}{\sigma_2}\frac{\partial M_2}{\partial C_6}\Delta C_6\\ \cdot\\ \cdot\\ \cdot\\ \frac{1}{\sigma_n}\frac{\partial M_n}{\partial C_1}\Delta C_1+\frac{1}{\sigma_n}\frac{\partial M_n}{\partial C_2}\Delta C_2+\cdots+\frac{1}{\sigma_n}\frac{\partial M_n}{\partial C_6}\Delta C_6\end{bmatrix}=\begin{bmatrix}\frac{\delta_1}{\sigma_1}\\ \frac{\delta_2}{\sigma_2}\\ \cdot\\ \cdot\\ \cdot\\ \frac{\delta_n}{\sigma_n}\end{bmatrix}$$

Rearranging the middle column matrix, the above equation can be written as

$$\begin{bmatrix}\frac{\Delta m_1}{\sigma_1}\\ \frac{\Delta m_2}{\sigma_2}\\ \cdot\\ \cdot\\ \cdot\\ \frac{\Delta m_n}{\sigma_n}\end{bmatrix}-\begin{bmatrix}\frac{1}{\sigma_1}\frac{\partial M_1}{\partial C_1}\frac{1}{\sigma_1}\frac{\partial M_1}{\partial C_2}\cdots\frac{1}{\sigma_1}\frac{\partial M_1}{\partial C_6}\\ \cdot\quad\cdot\quad\cdot\\ \cdot\quad\cdot\quad\cdot\\ \cdot\quad\cdot\quad\cdot\\ \frac{1}{\sigma_n}\frac{\partial M_n}{\partial C_1}\frac{1}{\sigma_n}\frac{\partial M_n}{\partial C_2}\cdots\frac{1}{\sigma_n}\frac{\partial M_n}{\partial C_6}\end{bmatrix}\begin{bmatrix}\Delta C_1\\ \Delta C_2\\ \cdot\\ \cdot\\ \Delta C_6\end{bmatrix}=\begin{bmatrix}\frac{\delta_1}{\sigma_1}\\ \frac{\delta_2}{\sigma_2}\\ \cdot\\ \cdot\\ \cdot\\ \frac{\delta_n}{\sigma_n}\end{bmatrix}$$

or equivalently written in shorthand as

$$\Delta m-\Lambda\Delta C=\delta \qquad (7)$$

Using the same notation, the weighted least square criteria established in part A as

$$(\partial/\partial\Delta C_j)\sum_{i=1}^{n}(\delta_i/\sigma_i)^2=0$$

is written in matrix notation as

$$(\partial/\partial \Delta C)\delta^T\delta = 0 \tag{8}$$

where

$$\delta^T\delta = \Delta m^T \Delta m - \Delta m^T \Lambda \Delta C - \Delta C^T \Lambda^T \Delta m + \Delta C^T \Lambda^T \Lambda \Delta C \tag{9}$$

and T denotes the transpose of the respective matrices. Performing the partials with respect to ΔC indicated by equation 8 gives the best estimate of the orbital parameters in terms of the residuals, or observational minus computed differences. That is,

$$\Delta C = (\Lambda^T \Lambda)^{-1} \Lambda^T \Delta m \tag{10}$$

The variance-covariance matrix of the orbital parameter corrections is

$$Q \equiv E(\Delta C \Delta C^T) = E\{(\Lambda^T\Lambda)^{-1}\Lambda^T \Delta m [(\Lambda^T\Lambda)^{-1}\Lambda^T \Delta m]^T\} \tag{11}$$

It is assumed that there is no correlation between observations (i.e., the covariances are zero); hence

$$E(\Delta m \Delta m^T) = I \tag{12}$$

and equation 11 becomes

$$Q = (\Lambda^T \Lambda)^{-1} \tag{13}$$

The Q matrix is called the variance–covariance matrix of the orbital elements. This matrix gives an estimate of the quality of prediction of the orbital elements for a given tracking schedule. The square root of the diagonal elements of this matrix are the standard deviations of the orbital elements. This matrix is calculated from two-body theory and need only be calculated once for a given orbit and set of data. The major objective of the following sections of this document is to show how Q is obtained.

EQUATIONS OF THE ORBIT PREDICTION ACCURACY PROGRAM

Satellite Position and Velocity Calculations

The calculation of the position and velocity of the satellite at any time t in the planar polar coordinate system (r, θ, $\dot{r}$, $r\dot{\theta}$), in the orbital plane system (X, Y, Z, $\dot{X}$, $\dot{Y}$, $\dot{Z}$), in the tracking system (x_s, y_s, z_s, $\dot{x}_s$, $\dot{y}_s$, $\dot{z}_s$) given the orbital elements a, e, i, ω, Ω, t_p is as follows: (See appendix A for discussion of the coordinate systems).

Planar Polar Coordinate System

The position and velocity of the satellite for any time t in planar polar coordinates is the solution of the following equations:

$$n = (\mu)^{1/2} a^{-3/2} \tag{14}$$

$$M = n(t - t_p) \tag{15}$$

$$M = E - e \sin E \tag{16}$$

$$r = a(1 - e \cos E) \tag{17}$$

$$\theta = 2 \tan^{-1}\{[(1+e)/(1-e)]^{\frac{1}{2}} \tan E/2\} \tag{18}$$

$$\phi = \theta + \omega \tag{19}$$

$$\dot{r} = [(nae \sin E)/(1 - e \cos E)] \tag{20}$$

$$\dot{\theta} = \dot{\phi} = \{[\mu a(1-e^2)]^{1/2}/r^2\} \tag{21}$$

where the determination of $E = E(M)$, equation 16 involves an iteration procedure. The process is as follows; assuming an estimate of the solution as E_0, equation 16 can be expanded in a Taylor series about E_0[3]. Keeping linear terms, then

$$M = E_0 - e \sin E_0 + (1 - e \cos E_0)\Delta E_0$$

or

$$\Delta E_0 = \frac{M - E_0 + e \sin E_0}{1 - e \cos E_0}$$

Taking $E_1 = E_0 + \Delta E_0$, the process may be repeated until there is no significant change in E_1 from one calculation to another.

Station Fixed Coordinate System

The coordinates of the satellites in station space are

$$\begin{bmatrix} x_s \\ y_s \\ z_s \end{bmatrix} = B \begin{bmatrix} r \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} R \\ 0 \\ 0 \end{bmatrix} \tag{22}$$

The velocity components of the satellite with

respect to the station are

$$\begin{bmatrix} \dot{x}_s \\ \dot{y}_s \\ \dot{z}_s \end{bmatrix} = B \begin{bmatrix} \dot{r} \\ r\dot{\theta} \\ 0 \end{bmatrix} + \dot{S}\Theta^T \begin{bmatrix} r \\ 0 \\ 0 \end{bmatrix} \tag{23}$$

where the matrices B, S and Θ^T are defined in appendix A.

The slant range to the satellite is

$$\rho = (x_s^2 + y_s^2 + z_s^2)^{1/2} \tag{24}$$

The direction cosines are

$$\begin{aligned} n &= \cos\alpha = x_s/\rho \\ l &= \cos\beta = y_s/\rho \\ m &= \cos\gamma = z_s/\rho \end{aligned} \tag{25}$$

The elevation angle is

$$E = \tfrac{1}{2}\pi - \alpha \tag{26}$$

The azimuth is given by

$$\begin{aligned} \cos A &= y_s/(y_s^2 + z_s^2)^{1/2}, \\ \sin A &= -z_s/(y_s^2 + z_s^2)^{1/2} \end{aligned} \tag{27}$$

The slant range velocity of the satellite is given by

$$\dot{\rho} = (\dot{x}_s^2 + \dot{y}_s^2 + \dot{z}_s^2)^{1/2} \tag{28}$$

The radial component of the slant range velocity (referred to as slant range-rate) is given by

$$\dot{\rho}_R = \frac{\dot{\bar{\rho}} \cdot \bar{\rho}}{\rho} = \frac{x_s\dot{x}_s + y_s\dot{y}_s + z_s\dot{z}_s}{\rho} \tag{29}$$

The tracking station location with respect to inertial space at any time t is found by

$$\lambda = \omega_e(t - t_0) + \lambda_s \tag{30}$$

where ω_e = rotational velocity of the earth ($7.29211585 \times 10^{-5}$ rads/sec).

Orbital Plane Coordinate System

The position of the satellite in the orbital plane coordinate system is given by

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \Theta \begin{bmatrix} x \\ y \\ z \end{bmatrix} = B^{-1} \begin{bmatrix} x_s + R \\ y_s \\ z_s \end{bmatrix} \tag{31}$$

where B and Θ are defined in appendix A.

The velocity of the satellite in the orbital plane coordinate system is given by

$$\begin{bmatrix} \dot{X}' \\ \dot{Y}' \\ \dot{z}' \end{bmatrix} = \begin{bmatrix} \dot{r} \\ r\dot{\theta} \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ r\dot{\theta} \\ -0 \end{bmatrix} \tag{32}$$

where $(0, r\dot{\theta}, 0)$ is the velocity of the (X, Y, Z) system with respect to the (x, y, z) system at latitude, L.

Define from equation 32 the nominal velocity as;

$$\begin{bmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{bmatrix} \equiv \begin{bmatrix} \dot{r} \\ r\dot{\theta} \\ 0 \end{bmatrix} \tag{33}$$

Prediction Accuracy of the Orbital Elements

The weighted least squares estimate of the variance-covariance matrix of the orbital parameters was derived in section III–B, equation 13, assuming no correlation between any observations. This relation is

$$Q = (\Lambda^T\Lambda)^{-1} \tag{34}$$

where

$$\Lambda = \begin{bmatrix} \frac{1}{\sigma_1}\frac{\partial M_1}{\partial C_1} & \frac{1}{\sigma_1}\frac{\partial M_1}{\partial C_2} & \cdots & \frac{1}{\sigma_1}\frac{\partial M_1}{\partial C_6} \\ \frac{1}{\sigma_2}\frac{\partial M_2}{\partial C_1} & \frac{1}{\sigma_2}\frac{\partial M_2}{\partial C_2} & \cdots & \frac{1}{\sigma_2}\frac{\partial M_2}{\partial C_6} \\ \vdots & & & \vdots \\ \frac{1}{\sigma_n}\frac{\partial M_n}{\partial C_1} & \frac{1}{\sigma_n}\frac{\partial M_n}{\partial C_2} & \cdots & \frac{1}{\sigma_n}\frac{\partial M_n}{\partial C_6} \end{bmatrix} \tag{35}$$

The σ_i is the known estimate of the standard deviation for the ith measurement. $\partial M_i/\partial C_j$ is the partial derivative of the ith measurement with respect to the jth orbit parameter.

These partials $\partial M_i/\partial C_j$ are computed in closed form using the classical two-body orbit theory. The assumption of this model is made since only a few significant figures are required in the elements of Q.

The actual calculations of the partials in the Λ matrix are done by two straight forward calculations. That is, consider the ith measurement, or the ith row of the Λ matrix as

$$N=\sigma_i^{-1}(\partial M_i/\partial C_i) \quad \text{where } j=1\cdots 6 \tag{36}$$

Expanding the partials by the chain rule results in

$$N=\frac{1}{\sigma_i}\left[\frac{\partial M_i}{\partial X}\frac{\partial X}{\partial C_j}+\frac{\partial M_i}{\partial Y}\frac{\partial Y}{\partial C_j}+\frac{\partial M_i}{\partial Z}\frac{\partial Z}{\partial C_j}+\frac{\partial M_i}{\partial \dot{X}}\frac{\partial \dot{X}}{\partial C_j}+\frac{\partial M_i}{\partial \dot{Y}}\frac{\partial \dot{Y}}{\partial C_j}+\frac{\partial M_i}{\partial \dot{Z}}\frac{\partial \dot{z}}{\partial C_j}\right] \tag{37}$$

where $j=1\cdots 6$

Define the row matrix

$$F(X,\ Y,\ Z)\equiv\left(\frac{\partial M_i}{\partial X},\ \frac{\partial M_i}{\partial Y},\ \frac{\partial M_i}{\partial Z}\right) \tag{38}$$

$$F(\dot{X},\ \dot{Y},\ \dot{Z})\equiv\left(\frac{\partial M_i}{\partial \dot{X}},\ \frac{\partial M_i}{\partial \dot{Y}},\ \frac{\partial M_i}{\partial \dot{Z}}\right) \tag{39}$$

and defining the 3×6 matrices as

$$P_{\text{pos}}\equiv\left(\frac{\partial X}{\partial C_j},\ \frac{\partial Y}{\partial C_j},\ \frac{\partial Z}{\partial C_j}\right)^T \quad \text{where } j=1\cdots 6 \tag{40}$$

$$P_{\text{vel}}\equiv\left(\frac{\partial \dot{X}}{\partial C_j},\ \frac{\partial \dot{Y}}{\partial C_j},\ \frac{\partial \dot{Z}}{\partial C_j}\right)^T \tag{41}$$

The ith row in the matrix of partials with respect to the orbital parameters can be written symbolically as

$$N=\sigma_i^{-1}[F(X,\ Y,\ Z)P_{\text{pos}}+F(\dot{X},\ \dot{Y},\ \dot{Z})P_{\text{vel}}] \tag{42}$$

If the observation is dependent only upon position, then

$$F(\dot{X},\ \dot{Y},\ \dot{Z})\equiv 0.$$

This is true for calculations pertaining to range, azimuth, elevation and the minitrack direction cosines.

For the range-rate observations;

$$F(\dot{X},\ \dot{Y},\ \dot{Z})\neq 0.$$

In other words, the partials of range-rate are dependent upon both position and velocity. (See equation 29.)

The matrices $F(X,\ Y,\ Z)$, $F(\dot{X},\ \dot{Y},\ \dot{Z})$, P_{pos} and P_{vel} contain partial derivatives and are evaluated as follows:

Evaluation of F(X, Y, Z) and F(Ẋ, Ẏ, Ż)

These row matrices are given by

$$F(X,\ Y,\ Z)=F(x_s,\ y_s,\ z_s)B+F(\dot{x}_s,\ \dot{y}_s,\ \dot{z}_s)\dot{S}\Theta^T \tag{43}$$

and

$$F(\dot{X},\ \dot{Y},\ \dot{Z})=F(\dot{x}_s,\ \dot{y}_s\ \ \dot{z}_s)B \tag{44}$$

where

$$F(x_s,\ y_s,\ z_s)\equiv\left[\frac{\partial M_i}{\partial x_s},\ \frac{\partial M_i}{\partial y_s},\ \frac{\partial M_i}{\partial z_s}\right] \tag{45}$$

$$F(\dot{x}_s,\ \dot{y}_s,\ \dot{z}_s)\equiv\left[\frac{\partial M_i}{\partial \dot{x}_s},\ \frac{\partial M_i}{\partial \dot{y}_s},\ \frac{\partial M_i}{\partial \dot{z}_s}\right] \tag{46}$$

B is a matrix which is defined in appendix A–3. The matrices $F(x_s,\ y_s,\ z_s)$ and $F(\dot{x}_s,\ \dot{y}_s,\ \dot{z}_s)$ are presented in appendix B for different types of observations. The matrices S and Θ are defined in appendix A.

The proof of equations 43 and 44 is as follows: Consider the ith observable M_i

$$\Delta M_i=\frac{\partial M_i}{\partial x_s}\Delta x_s+\frac{\partial M_i}{\partial y_s}\Delta y_s+\frac{\partial M_i}{\partial z_s}\Delta z_s+\frac{\partial M_i}{\partial \dot{x}_s}\Delta \dot{x}_s+\frac{\partial M_i}{\partial \dot{y}_s}\Delta \dot{y}_s+\frac{\partial M_i}{\partial \dot{z}_s}\Delta \dot{z}_s \tag{47}$$

From appendix A–3

$$\begin{bmatrix}\Delta x_s\\ \Delta y_s\\ \Delta z_s\end{bmatrix}=B\begin{bmatrix}\Delta X\\ \Delta Y\\ \Delta Z\end{bmatrix} \tag{48}$$

and

$$\begin{bmatrix}\Delta\dot{x}_s\\ \Delta\dot{y}_s\\ \Delta\dot{z}_s\end{bmatrix}=B\begin{bmatrix}\Delta\dot{X}\\ \Delta\dot{Y}\\ \Delta\dot{Z}\end{bmatrix}+\dot{S}\Theta^T\begin{bmatrix}\Delta X\\ \Delta Y\\ \Delta Z\end{bmatrix} \tag{49}$$

Substitution of equations 48 and 49 into equation 47 gives the result that

$$\Delta M_i=\left(\frac{\partial M_i}{\partial x_s},\frac{\partial M_i}{\partial y_s},\frac{\partial M_i}{\partial z_s}\right)B[\,\Delta X\Delta Y\Delta Z]^T$$

$$+\left(\frac{\partial M_i}{\partial x_s},\frac{\partial M_i}{\partial y_s},\frac{\partial M_i}{\partial z_s}\right)$$

$$\times\{B[\,\Delta\dot{X}\Delta\dot{Y}\Delta\dot{Z}]^T+\dot{S}\Theta^T[\,\Delta X\Delta Y\Delta Z]^T\} \tag{50}$$

Or, using the aforementioned notation, equation 50 is written as;

$$\Delta M_i=\{F(x_s,\,y_s,\,z_s)B+F(\dot{x}_s,\,\dot{y}_s,\,\dot{z}_s)\dot{S}\Theta^T\}$$

$$\times[\,\Delta X\Delta Y\Delta Z]^T+F(\dot{x}_s,\,\dot{y}_s,\,\dot{z}_s)B[\,\Delta\dot{X}\Delta\dot{Y}\Delta\dot{Z}]^T \tag{51}$$

By definition

$$\Delta M_i=F(X,\,Y,\,Z)\begin{bmatrix}\Delta X\\ \Delta Y\\ \Delta Z\end{bmatrix}+F(\dot{X},\,\dot{Y},\,\dot{Z})\begin{bmatrix}\Delta\dot{X}\\ \Delta\dot{Y}\\ \Delta\dot{Z}\end{bmatrix} \tag{52}$$

Inspection of equations 51 and 52 leads to the result that

$$F(X,\,Y,\,Z)=F(x_s,\,y_s,\,z_s)B+F(\dot{x}_s,\,\dot{y}_s,\,\dot{z}_s)\dot{S}\Theta^T \tag{53}$$

and

$$F(\dot{X},\,\dot{Y},\,\dot{Z})=F(\dot{x}_s,\,\dot{y}_s,\,\dot{z}_s)B \tag{54}$$

Evaluation of P_{pos} and P_{vel}

The vectors $(\Delta X,\,\Delta Y,\,\Delta Z)^T$ and $(\Delta\dot{X},\,\Delta\dot{Y},\,\Delta\dot{Z})^T$ are related to the vector ΔC by the relationships

$$\begin{bmatrix}\Delta X\\ \Delta Y\\ \Delta Z\end{bmatrix}=P_{pos}\Delta C \tag{55}$$

and

$$\begin{bmatrix}\Delta\dot{X}\\ \Delta\dot{Y}\\ \Delta\dot{Z}\end{bmatrix}=P_{vel}\Delta C \tag{56}$$

where P_{pos} is a matrix which contains the partials of the position coordinates X, Y, Z with respect to the orbital parameters. The elements of P_{pos} are given in appendix C–1; and where P_{vel} is a matrix which contains the partials of the velocity components $\dot{X}$, $\dot{Y}$, $\dot{Z}$ with respect to the orbital parameters. The elements are given in appendix C–2.

Confidence Ellipsoid of the Position and Velocity Coordinates

The variance-covariance matrix of the orbital elements, Q, is calculated by the method presented in part B of this section. The square root of the diagonal of this 6×6 matrix represents the standard deviation of the orbital elements for a given tracking schedule and orbit. It is of interest to investigate how the estimate of the errors in the orbital elements are propagated in position and velocity at various points in the orbit.

The variance–covariance matrices of position and velocity at a given time are referred to as position and velocity confidence ellipsoids. These ellipsoids are the boundaries of a region in which the satellite is located (in position and velocity space) with a certain probability. That is, the probability that an error in the satellite's coordinate will be within the calculated standard deviation is equal to the probability that the value of the coordinate falls within that standard deviation. For example, if the calculated standard deviation in the X-coordinate is 20m, the probability that the satellite lies within 20m is 68.2%; 40m is 95.4%; 60 m is 99.72%. The calculation of the variance-covariance matrix of position and velocity at different times in the orbit is done by the following:

Position Confidence Ellipsoid

The incremental change in the position vector in the orbiting coordinate system at any time is

obtained by

$$\begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} = P_{\text{pos}}(t)\Delta C \qquad (57)$$

The variance–covariance matrix of the position in this system is given by

$$U_{\text{pos}}(t) = P_{\text{pos}} Q P^T{}_{\text{pos}} \qquad (58)$$

where $U_{\text{pos}}(t)$ is a 3×3 matrix

Velocity Confidence Ellipsoid

The incremental change in the velocity vector in the orbiting coordinate system at any time t is obtained by

$$\begin{bmatrix} \Delta \dot{X} \\ \Delta \dot{Y} \\ \Delta \dot{Z} \end{bmatrix} = P_{\text{vel}}(t)\Delta C \qquad (59)$$

The variance-covariance matrix of the velocity in this system is given by

$$U_{\text{vel}}(t) = P_{\text{vel}} Q P^T{}_{\text{vel}} \qquad (60)$$

where $U_{\text{vel}}(t)$ is a 3×3 matrix. The "sandwich" transformations of equations 58 and 60 can be shown as follows: by operating both sides of equation 57 by the transpose results in

$$\Delta X \Delta X^T = P_{\text{pos}} \Delta C (P_{\text{pos}} \Delta C)^T$$

$$= P_{\text{pos}} \Delta C \Delta C^T P^T{}_{\text{pos}}$$

taking the expected value of the above equation gives

$$U_{\text{pos}} \equiv E(\Delta X \Delta X^T) = P_{\text{pos}} Q P^T{}_{\text{pos}}$$

The same operation is done for obtaining U_{vel}.

The variance-covariance matrix calculated in the X, Y, Z coordinates normally have cross-product terms. The 3×3 matrices are diagonalized to find the principal values (or eigenvalues) of the variances. The eigenvectors, which define the direction of the principal axis with respect to the X, Y, Z axis, can be determined by the following[4]:

For the 3×3 matrix, U_{pos}, the roots of the equation

$$| U_{\text{pos}} - I\lambda | = 0 \qquad (61)$$

are found, giving 3 values of λ, called the eigenvalues of the U_{pos} matrix. Defining the eigenvectors of λ_i (where $i = 1, 2, 3$) as $\bar{R}_i$, the relationship for finding the components of the eigenvectors is

$$(U_{\text{pos}} - I\lambda_i)\bar{R}_i = 0 \qquad (62)$$

where, for a given λ_i the 3 equations are solved simultaneously for the components of $\bar{R}_i$. Similar procedures can be used to find the eigenvalues and eigenvectors of U_{vel}.

REFERENCES

1. Kraft, J. D., and Salinger, R. G., "Special Purpose Imp Orbits which Emphasize the Moon's Gravitational Effects," June 1963, Contract No. NAS 5-1611, Task Order No. 11.
2. Leach, R., "Evaluating the Quality of Prediction for a Position-Prediction or Tracking System," ARS Journal, November 1962.
3. Martin Personnel, Design Guide to Orbital Flight.
4. Goldstein, H., *Classical Mechanics*, Addison-Wesley Publ. Co., copyright 1959.

Appendix A

COORDINATE SYSTEMS

A-1 TRANSFORMATION FROM THE INERTIAL SYSTEM (x, y, z) TO THE ORBITAL PLANE SYSTEM (X, Y, Z)

The coordinates x, y, z have their origin at the center of the Earth, the positive x-axis in the direction of the vernal equinox, the positive z-axis along the north pole of the earth, and the positive y-axis forms a right-hand set.

The coordinates X, Y, Z have their origin at the center of the earth, the positive X-axis in the direction of the radius vector from the center of the earth to the satellite, the positive Z-axis is in the direction of the angular momentum vector and the positive Y-axis lies in the orbital plane and forms a right-hand set.

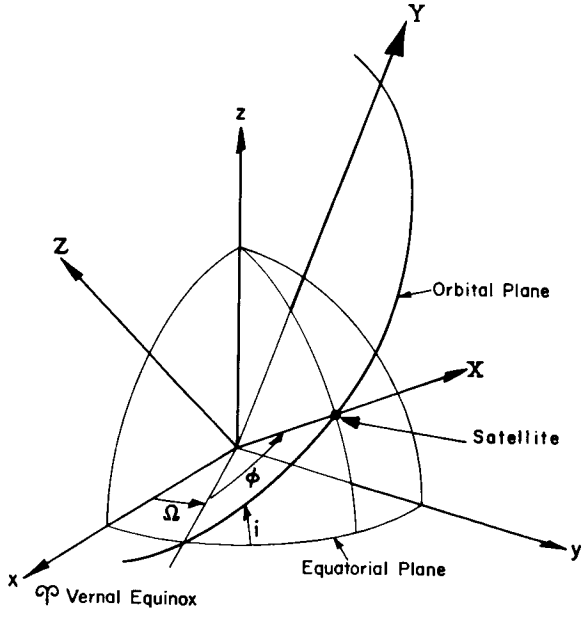


FIGURE A–1.—Schematic of the Inertial Coordinates (x, y, z) and the Orbiting Plane Coordinates (X, Y, Z).

These two coordinate systems are shown in figure A–1 and are related by the transformation

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \Theta \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{A-1}$$

where

$$\Theta \equiv \begin{bmatrix} c\phi c\Omega - s\phi cis\Omega & c\phi s\Omega + s\phi cic\Omega & s\phi si \\ -s\phi c\Omega - c\phi cis\Omega & (-s\phi s\Omega + c\phi cic\Omega) & c\phi si \\ sis\Omega & -sic\Omega & ci \end{bmatrix}$$

The derivation of the Θ matrix is from the following rotations

$$\Theta = A_3(\phi) A_1(i) A_3(\Omega)$$

A-2 TRANSFORMATION FROM THE INERTIAL COORDINATE SYSTEM (x, y, z) TO THE STATION COORDINATE SYSTEM (x_s, y_s, z_s)

The coordinates x, y, z are defined in part A–1 above.

The station coordinates have their origin at the station (assumed to be on the reference spheroid), the x_s-axis points outward normal to the reference spheroid, the y_s-axis is tangent to the reference spheroid and it points north, the z_s-axis forms a right hand set and it points west.

These two coordinate systems are shown in figure A–2 and are related by

$$\begin{bmatrix} x_s + R \\ y_s \\ z_s \end{bmatrix} = S \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{A-2}$$

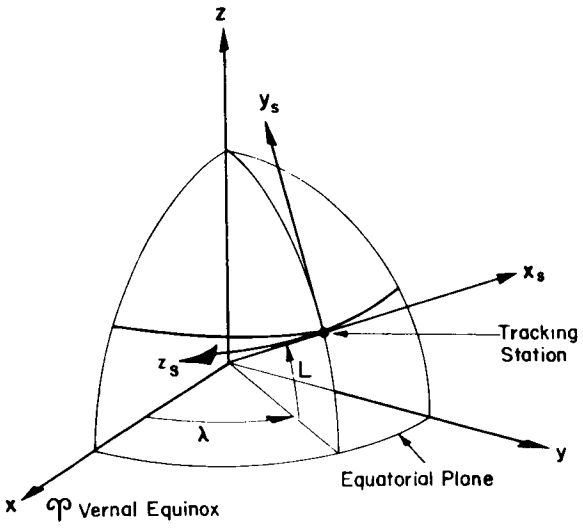


FIGURE A–2.—Schematic of the Inertial Coordinates (x, y, z) and the Tracking Station Coordinates (x_s, y_s, z_s).

where $R=$ the equatorial radius of the earth (6378.165 km)

$$S \equiv \begin{bmatrix} cLc\lambda & cLs\lambda & sL \\ -sLc\lambda & -sLs\lambda & cL \\ s\lambda & -c\lambda & 0 \end{bmatrix}$$

A-3 TRANSFORMATION FROM THE ORBITAL PLANE COORDINATE SYSTEM (X, Y, Z) TO THE STATION COORDINATE SYSTEM (x_s, y_s, z_s)

These two coordinate systems are defined in parts A–1 and A–2. They are related by the transformation

$$\begin{bmatrix} x_s+R \\ y_s \\ z_s \end{bmatrix} = B \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

where

$$B = S\Theta^T$$

The matrices Θ and S are defined in A–1 and A–2 above, respectively.

For two body theory the satellite travels in the XY-plane and $Y=Z=0$, hence

$$\begin{bmatrix} x_s+R \\ y_s \\ z_s \end{bmatrix} = B \begin{bmatrix} r \\ 0 \\ 0 \end{bmatrix}$$

Appendix B

THE PARTIAL DERIVATIVES OF THE OBSERVABLES WITH RESPECT TO THE STATION COORDINATES

B-1: RANGE, ρ

The matrix of the partial derivatives for the range is given by

$$F(x_s, y_s, z_s) = \left[\frac{\partial \rho}{\partial x_s}, \frac{\partial \rho}{\partial y_s}, \frac{\partial \rho}{\partial z_s}\right]$$

differentiation of equation 24 give

$$F(x_s, y_s, z_s) = \left[\frac{x_s}{\rho}, \frac{y_s}{\rho}, \frac{z_s}{\rho}\right]$$

from equations 25

$$F(x_s, y_s, z_s) = [n, l\ m]$$

B-2: RANGE-RATE, $\dot{\rho}_R$

The range-rate partials are given by the following two matrices

$$F(x_s, y_s, z_s) = \left(\frac{\partial \dot{\rho}_R}{\partial x_s}, \frac{\partial \dot{\rho}_R}{\partial y_s}, \frac{\partial \dot{\rho}_R}{\partial z_s}\right)$$

$$F(\dot{x}_s, \dot{y}_s, \dot{z}_s) = \left(\frac{\partial \dot{\rho}_R}{\partial \dot{x}_s}, \frac{\partial \dot{\rho}_R}{\partial \dot{y}_s}, \frac{\partial \dot{\rho}_R}{\partial \dot{z}_s}\right)$$

By performing the indicated differentiation of equation 29, the range-rate partials become:

$$F(x_s, y_s, z_s) = \left[\frac{\dot{x}_s}{\rho} - Dx_s, \frac{\dot{y}_s}{\rho} - Dy_s, \frac{\dot{z}_s}{\rho} - Dz_s\right]$$

$$F(\dot{x}_s, \dot{y}_s, \dot{z}_s) = \left[\frac{x_s}{\rho}, \frac{y_s}{\rho}, \frac{z_s}{\rho}\right]$$

where

$$D \equiv [(x_s\dot{x}_s + y_s\dot{y}_s + z_s\dot{z}_s)/\rho^3]$$

The values of $F(x_s, y_s, z_s)$ and $F(\dot{x}_s, \dot{y}_s, \dot{z}_s)$ computed from these equations are substituted into equations 43 and 44 to give $F(X, Y, Z)$ and $F(\dot{X}, \dot{Y}, \dot{Z})$, respectively. Subsequently, $F(X, Y, Z)$ and $F(\dot{X}, \dot{Y}, \dot{Z})$ are substituted into equation 42 along with P_{pos} and P_{vel} (from appendix C) to give N. The matrix N is the ith row matrix of the partials of the range-rate with respect to the orbit parameters. Symbolically, the row matrix of Λ is written as;

$$N = \frac{1}{\sigma_{\dot{\rho}R}}\left[\frac{\partial \dot{\rho}_R}{\partial a}, \frac{\partial \dot{\rho}_R}{\partial e}, \frac{\partial \dot{\rho}_R}{\partial t_p}, \frac{\partial \dot{\rho}_R}{\partial \omega}, \frac{\partial \dot{\rho}_R}{\partial \Omega}, \frac{\partial \dot{\rho}_R}{\partial i}\right]$$

B-3: AZIMUTH, A

The matrix of the partial derivatives of the azimuth is given by

$$F(x_s, y_s, z_s) = \left[\frac{\partial A}{\partial x_s}, \frac{\partial A}{\partial y_s}, \frac{\partial A}{\partial z_s}\right]$$

Differentiation of equation 27 gives

$$F(x_s, y_s, z_s) = \left[0, \frac{z_s}{y_s^2 + z_s^2}, \frac{-y_s}{y_s^2 + z_s^2}\right]$$

The value of $F(x_s, y_s, z_s)$ obtained from this equation are substituted into equation (42) along with P_{pos} (from app. C) to give N. The matrix N is the row matrix of the partials of the azimuth with respect to the orbit parameters.

B-4: ELEVATION, E

The matrix of the partial derivatives of the elevation is given by

$$F(x_s, y_s, z_s) = \left[\frac{\partial E}{\partial x_s}, \frac{\partial E}{\partial y_s}, \frac{\partial E}{\partial z_s}\right]$$

Differentiating equation 26 gives

$$F(x_s, y_s, z_s) = \left[\frac{\cos E}{\rho}, \frac{-\sin E \cos A}{\rho}, \frac{\sin E \sin A}{\rho}\right]$$

The value of $F(x_s, y_s, z_s)$ obtained from this equation is used to find N, a row of the Λ matrix.

B-5: DIRECTION COSINES, *n, l, m*

The matrices of the partial derivatives of the direction cosines are given by the following:

For n:

$$F(x_s, y_s, z_s) = \left[\frac{\partial n}{\partial x_s}, \frac{\partial n}{\partial y_s}, \frac{\partial n}{\partial z_s}\right]$$

Differentiating equation 25 gives

$$F(x_s, y_s, z_s) = \left[\frac{1-n^2}{\rho}, \frac{-ln}{\rho}, \frac{-mn}{\rho}\right]$$

For l:

$$F(x_s, y_s, z_s) = \left[\frac{\partial l}{\partial x_s}, \frac{\partial l}{\partial y_s}, \frac{\partial l}{\partial z_s}\right]$$

Differentiating equation 25 gives

$$F(x_s, y_s, z_s) = \left[\frac{-ln}{\rho}, \frac{1-l^2}{\rho}, \frac{-lm}{\rho}\right]$$

For m:

$$F(x_s, y_s, z_s) = \left[\frac{\partial m}{\partial x_s}, \frac{\partial m}{\partial y_s}, \frac{\partial m}{\partial z_s}\right]$$

Differentiating equation 25 gives

$$F(x_s, y_s, z_s) = \left[\frac{-mn}{\rho}, \frac{-lm}{\rho}, \frac{1-m^2}{\rho}\right]$$

Each direction cosine gives a row of Λ matrix

Appendix C

THE PARTIAL DERIVATIVES OF THE ORBITAL PLANE COORDINATES WITH RESPECT TO THE ORBITAL PARAMETERS

C-1: PM7 MATRIX

The P_{pos} matrix relates changes in the orbital elements to corresponding changes in position by

$$\begin{bmatrix}\Delta X\\ \Delta Y\\ \Delta Z\end{bmatrix} = P_{pos}\begin{bmatrix}\Delta a\\ \Delta e\\ \Delta t_p\\ \Delta\omega\\ \Delta\Omega\\ \Delta i\end{bmatrix} \qquad \text{(C-1)}$$

The relationship between the (X, Y, Z) coordinate system and the (x, y, z) coordinate system is given by (see appendix A-1).

$$\begin{bmatrix}X\\ Y\\ Z\end{bmatrix} = \Theta\begin{bmatrix}x\\ y\\ z\end{bmatrix} \qquad \text{(C-2)}$$

The matrix Θ transforms a vector in (x, y, z) system to a vector in (X, Y, Z) system. The vector (dx, dy, dz) transformed by the same matrix Θ so that

$$\begin{bmatrix}\Delta X\\ \Delta Y\\ \Delta Z\end{bmatrix} = \Theta\begin{bmatrix}\Delta x\\ \Delta y\\ \Delta z\end{bmatrix} \qquad \text{(C-3)}$$

the relationships between the inertial system (x, y, z) and the satellites position in terms of orbital elements is given by

$$\begin{bmatrix}x\\ y\\ z\end{bmatrix} = \Theta^T\begin{bmatrix}r\\ 0\\ 0\end{bmatrix} \qquad \text{(C-4)}$$

differentiation of equations C-4 written symbolically as

$$\begin{bmatrix}\Delta x\\ \Delta y\\ \Delta z\end{bmatrix} = \Delta(\Theta^T)\begin{bmatrix}r\\ 0\\ 0\end{bmatrix} + \Theta^T\begin{bmatrix}\Delta r\\ 0\\ 0\end{bmatrix} \qquad \text{(C-5)}$$

can be substituted into equation C-3 to obtain

$$\begin{bmatrix}\Delta X\\ \Delta Y\\ \Delta Z\end{bmatrix} = \Theta\Delta(\Theta^T)\begin{bmatrix}r\\ 0\\ 0\end{bmatrix} + \begin{bmatrix}\Delta r\\ 0\\ 0\end{bmatrix} \qquad \text{(C-6)}$$

Performing the operations indicated by equation C-6 gives

$$\begin{bmatrix}\Delta X\\ \Delta Y\\ \Delta Z\end{bmatrix} = \begin{bmatrix}\Delta r\\ r\Delta\phi + r\cos i\Delta\Omega\\ -r\sin i\cos\phi\Delta\Omega + r\sin\phi\Delta i\end{bmatrix} \qquad \text{(C-7)}$$

Total differentials are obtained by differentiating equations 14, 15, 16, 17, 18, and 19; and eliminating the differentials ΔM, Δn, ΔE, $\Delta\theta$. The results are;

$$\Delta r = \left[1 - e\cos E - \frac{3eM\sin E}{2(1-e\cos E)}\right]\Delta a + \frac{a(e-\cos E)}{1-e\cos E}\Delta e - \frac{e\sin E}{(1-e\cos E)}\left(\frac{\mu}{a}\right)^{1/2}\Delta t_p \qquad \text{(C-8)}$$

and

$$\Delta\phi = \Delta\omega + \frac{(1-e^2)^{1/2}}{(1-e\cos E)^2}\left\{-a^{-1}\left(\frac{\mu}{a}\right)^{1/2}\Delta t_p - \tfrac{3}{2}M\frac{\Delta a}{a} + \frac{\sin E}{(1-e^2)}(2 - e\cos E - e^2)\Delta e\right\} \qquad \text{(C-9)}$$

Equations C-8 and C-9 are substituted into

equation C–7 and the elements of the P_{pos} matrix are the partials or the coefficients of Δa, Δe, Δi, Δt_p, $\Delta\omega$, and $\Delta\Omega$. By inspection of the resulting equation, the element in the ith row and jth column of the P_{pos} matrix are as follows:

$$P_{11}=1-e\cos E-\frac{3eM\sin E}{2(1-e\cos E)}$$

$$P_{12}=\frac{(e-\cos E)a}{(1-e\cos E)}$$

$$P_{13}=\frac{-e\sin E}{(1-e\cos E)}\left(\frac{\mu}{a}\right)^{1/2}$$

$$P_{14}=P_{15}=P_{16}=0$$

$$P_{21}=\frac{-(1-e^2)^{1/2}}{(1-e\cos E)^2}\frac{3}{2}\frac{M}{a}r$$

$$P_{22}=\frac{(1-e^2)^{1/2}}{(1-e\cos E)^2}\frac{r}{(1-e^2)}\sin E(2-e\cos E-e^2)$$

$$P_{23}=\frac{-(1-e^2)^{1/2}}{(1-e\cos E)^2}\frac{r}{a}\left(\frac{\mu}{a}\right)^{1/2}$$

$$P_{24}=r$$

$$P_{25}=r\cos i$$

$$P_{26}=0$$

$$P_{31}=P_{32}=P_{33}=P_{34}=0$$

$$P_{35}=-r\cos\phi\sin i$$

$$P_{36}=r\sin\phi$$

C-2: P_{vel} MATRIX

The P_{vel} matrix relates the changes in the orbital elements to corresponding changes in velocity by

$$\begin{bmatrix}\Delta\dot X\\ \Delta\dot Y\\ \Delta\dot Z\end{bmatrix}=P_{vel}\begin{bmatrix}\Delta a\\ \Delta e\\ \Delta t_p\\ \Delta\omega\\ \Delta\Omega\\ \Delta i\end{bmatrix} \qquad \text{(C–10)}$$

The perturbation in $\dot x$, $\dot y$, and $\dot z$ are projected along the X, Y, Z axes by the transformation

$$\begin{bmatrix}\Delta\dot X\\ \Delta\dot Y\\ \Delta\dot Z\end{bmatrix}=\Theta\begin{bmatrix}\Delta\dot x\\ \Delta\dot y\\ \Delta\dot z\end{bmatrix} \qquad \text{(C–11)}$$

The inertial velocity of the satellite may be written as;

$$\begin{bmatrix}\dot x\\ \dot y\\ \dot z\end{bmatrix}=\Theta^T\begin{bmatrix}\dot r\\ r\dot\phi\\ 0\end{bmatrix} \qquad \text{(C–12)}$$

Differentiation of equation C–12 is written symbolically as;

$$\begin{bmatrix}\Delta\dot x\\ \Delta\dot y\\ \Delta\dot r\end{bmatrix}=\Delta(\Theta^T)\begin{bmatrix}\dot r\\ r\dot\phi\\ 0\end{bmatrix}+\Theta^T\begin{bmatrix}\Delta\dot r\\ \Delta(r\dot\phi)\\ 0\end{bmatrix} \qquad \text{(C–13)}$$

Substitution of equation C–13 into equation C–11 yields

$$\begin{bmatrix}\Delta\dot X\\ \Delta\dot Y\\ \Delta\dot Z\end{bmatrix}=\Theta\Delta(\Theta^T)\begin{bmatrix}\dot r\\ r\dot\phi\\ 0\end{bmatrix}+\begin{bmatrix}\Delta\dot r\\ \Delta(r\dot\phi)\\ 0\end{bmatrix} \qquad \text{(C–14)}$$

Performing the operations indicated by equation C–14 gives

$$\begin{bmatrix}\Delta\dot X\\ \Delta\dot Y\\ \Delta\dot Z\end{bmatrix}=\begin{bmatrix}\Delta\dot r-r\dot\phi\Delta\phi-r\dot\phi\cos i\Delta\Omega\\ \dot r\Delta\phi+r\Delta\dot\phi+\dot r\cos i\Delta\Omega+\dot\phi\Delta r\\ \sin i(r\dot\phi\sin\phi-\dot r\cos\phi)\Delta\Omega+(\dot r\sin\phi+r\dot\phi\cos\phi)\Delta i\end{bmatrix} \qquad \text{(C–15)}$$

The derivative of the eccentric anomaly, E, with respect to time is

$$\dot{E}=n/(1-e\cos E) \tag{C-16}$$

Differentiation of C-8 and C-9 with respect to time gives

$$\Delta\dot{r}=\frac{\dot{E}}{2(1-e\cos E)}\left\{-e\sin E(1-e\cos E)+\frac{3Me[e-\cos E]}{(1-e\cos E)}\right\}\Delta a+\frac{\sin E\dot{E}a}{(1-e\cos E)^2}(1-e^2)\Delta e$$

$$+\frac{e\dot{E}(e-\cos E)}{(1-e\cos E)^2}\left(\frac{\mu}{a}\right)^{1/2}\Delta t_p \tag{C-17}$$

and

$$\Delta\dot{\phi}=\frac{2e(1-e^2)^{1/2}(\sin E)\dot{E}}{a(1-e\cos E)^3}\left(\frac{\mu}{a}\right)^{1/2}\Delta t_p+\left\{\frac{6e(1-e^2)^{1/2}\dot{E}M(\sin E)}{(1-e\cos E)^3}-\frac{3(1-e^2)^{1/2}\dot{E}}{1-e\cos E}\right\}\frac{\Delta a}{2a} \tag{C-18}$$

$$+\frac{\dot{E}}{(1-e^2)^{1/2}(1-e\cos E)^3}[-3e+2\cos E+e^3(1+\sin^2 E)]\Delta e \tag{C-18}$$

Equations C-16, C-17, C-18, 20 and 21 are substituted into equation C-14 and the elements of the P_{vel} matrix are the coefficients of Δa, Δe, Δi, Δt_p, $\Delta\omega$, and $\Delta\Omega$. By inspection of the resulting equation, the elements in the ith row and jth column of the P_{vel} matrix are as follows:

$$P_{11}=\frac{n}{2(1-e\cos E)^2}\left[-e\sin E(1-e\cos E)+\frac{3Me(e-\cos E)}{1-e\cos E}\right]+\frac{3Mr\dot{\phi}(1-e^2)^{1/2}}{2a(1-e\cos E)^2}$$

$$P_{12}=\frac{n\sin E}{(1-e\cos E)^3}(1-e^2)a-\frac{r\dot{\phi}(1-e^2)^{1/2}\sin E(2-e\cos E-e^2)}{(1-e^2)(1-e\cos E)^2}$$

$$P_{13}=\frac{en(e-\cos E)}{(1-e\cos E)^3}\left(\frac{\mu}{a}\right)^{1/2}+\frac{r\dot{\phi}(1-e^2)^{1/2}}{(1-e\cos E)^2 a}\left(\frac{\mu}{a}\right)^{1/2}$$

$$P_{14}=-r\dot{\phi}$$

$$P_{15}=-r\dot{\phi}\cos i$$

$$P_{16}=0$$

$$P_{21}=-\frac{(1-e^2)^{1/2}}{(1-e\cos E)^2}\frac{3}{2}\frac{M\dot{r}}{a}+r\frac{3e(1-e^2)^{1/2}nM}{a(1-e\cos E)^4}\sin E-\frac{2(1-e^2)^{1/2}nr}{2a(1-e\cos E)^2}+\dot{\phi}\left[1-e\cos E-\frac{3eM\sin E}{2(1-e\cos E)}\right]$$

$$P_{22}=\frac{(1-e^2)^{1/2}\dot{r}}{(1-e\cos E)^2(1-e^2)}\sin E(2-e\cos E-e^2)+\frac{rn}{(1-e^2)^{1/2}(1-e\cos E)^4}$$

$$\times[-3e+2\cos E+e^3(1+\sin^2 E)]+\frac{\dot{\phi}a(e-\cos E)}{(1-e\cos E)}$$

$$P_{23}=-\frac{(1-e^2)^{1/2}\dot{r}}{(1-e\cos E)^2 a}\left(\frac{\mu}{a}\right)^{1/2}+\frac{2re(1-e^2)^{1/2}\sin En}{a(1-e\cos E)^4}\left(\frac{\mu}{a}\right)^{1/2}-\frac{\dot{\phi}e\sin E}{(1-e\cos E)}\left(\frac{\mu}{a}\right)^{1/2}$$

$$P_{24}=\dot{r}$$

$P_{25}=\dot{r}\cos i$

$P_{26}=0$

$P_{31}=P_{32}=P_{33}=P_{34}=0$

$P_{35}=-\dot{r}\cos\phi\sin i+r\sin i\sin\phi\dot{\phi}$

$P_{36}=\dot{r}\sin\phi+r\cos\phi\dot{\phi}$

ON DETERMINING THE SECULAR AND CRITICAL EFFECTS IN THE MOTION OF SATELLITES BY MEANS OF A NONSINGULAR SET OF VECTORIAL ELEMENTS

PETER MUSEN

Goddard Space Flight Center

Under the influence of the sun and the moon a highly eccentric orbit of a satellite can become a nearly circular one or a nearly circular orbit might become eccentric. In light of this I suggest a combination of the Gibbsian rotation vector and the Hamiltonian vector as a nonsingular set of elements to be used for all eccentricities and all inclinations. The method of Halphen is suggested for the computation of secular effects, and the method of Liouville for the computation of the near resonance effects caused by the commensurability of mean motions of the disturbed and the disturbing body. The combination of these two methods can give information about the general behavior of the orbit of a minor planet over an interval of some thousands of years, and of an artificial satellite moving in cislunar space over an interval of a number of years.

NOTATIONS

$\mathbf{r}$ = the position vector of the disturbed body.

$r = |\mathbf{r}|$.

ξ, η = the coordinates of the disturbed body in an ideal system of coordinates.

$M, \omega, \Omega, i, e, n$ = the standard elliptic elements of the disturbed body. The initial values of the elements are denoted by a zero subscript.

L = the mean orbital longitude.

ΔL = the perturbations in the mean orbital longitude.

$c^2 = p = a(1-e^2)$.

$\mathbf{P}, \mathbf{Q}, \mathbf{R}$ = the Gibbsian vectorial elements of the disturbed body. The elements of the disturbing body will be designated by the corresponding symbol with a prime.

χ = the true orbital longitude of the perigee reckoned from the departure point.

$u = -e\sin\chi/p^{1/2}$.

$v = e\cos\chi/p^{1/2}$.

E = the eccentric anomaly of the disturbed body.

$\lambda = E+\chi$, the eccentric orbital longitude.

$\mathbf{h}$ = the vector appearing in the Hamiltonian integral of the two-body problem.

$\mathbf{g}$ = the Gibbsian rotation vector.

$\mathbf{i}, \mathbf{j}, \mathbf{k}$ = the unit vectors along the axes of an ideal system rigidly connected to the osculating orbit plane.

$\mathbf{i}_0, \mathbf{j}_0, \mathbf{k}_0$ = the initial values of $\mathbf{i}, \mathbf{j}, \mathbf{k}$.

$\mathbf{F}$ = the disturbing force lying in the orbital plane of the disturbed body.

$\mathbf{F}_0$ = the disturbing force $\mathbf{F}$ averaged over the orbit of the disturbing body.

Z = the component of the disturbing force normal to the orbit plane of the disturbed body.

Z_0 = the component Z averaged over the orbit of the disturbing body.
X = the component of the disturbing force in the direction of $\mathbf{i}$.
Y = the component of the disturbing force in the direction of $\mathbf{j}$.
S = the component of the disturbing force in the direction of $\mathbf{r}$.
T = the component of the disturbing force in the direction of $\mathbf{k}\times\mathbf{r}$.

The gravitational constant is put equal to 1.

DISCUSSION

In a previous work I have emphasized the importance of long-period effects for the estimation of the lifetime of satellites moving in highly eccentric orbits (Musen, 1963). In that paper a collection of formulas was given for the computation of the long-period lunar effects in the motion of artificial satellites using Halphen's method (Halphen, 1888).

Originally I suggested the use of the classical elliptic elements, which are quite convenient if the eccentricity does not become small. The system of equations giving the variations of the elliptic elements contains the eccentricity and the sine of the inclination as "small divisors," thus making the computation very difficult if the orbit becomes a nearly circular one.

Unfortunately, this is exactly what may happen to originally highly eccentric orbits in cislunar space. The results obtained by Shute (1962) on the basis of Halphen's method show that under the influence of the sun and the moon a highly eccentric orbit can become a nearly circular one or a nearly circular orbit can become eccentric. Thus, the demarcation line between the two extreme types of orbits is not clearly defined.

In light of the accumulated information, I wish to reconsider the problem and to suggest a nonsingular set of elements to be used for all eccentricities and all inclinations.

This system consists of the vector appearing in the Hamiltonian expression for the velocity in the two-body problem and of the Gibbsian rotation vector (Gibbs, 1901). The Gibbsian vector determines the rotation of the orbital plane from its initial position to the position at a given time. I used the Hamiltonian vector (Musen, 1954) in the vectorial theory of the perturbations of elements. The Gibbsian vector was used (Musen, 1961) in an article on the generalization of Stromgren's method of special perturbations (Stromgren, 1929).

The combination of these two vectors represents a nonsingular system of elements, because no "small divisors" appear in the equations for their variations. Two sources of difficulty associated with the determination of perturbations for nearly circular orbits can be recognized. The first is our tendency to obtain the unit vector directed toward the perigee. This can be avoided by representing the position vector in terms of the Hamiltonian vector directly. The second source is the use of the eccentric and the mean anomalies. Both sources cause the presence of e as a small divisor.

The difficulty can be resolved by combining the perturbations in the perigee position with the mean and the eccentric anomalies, thus introducing the eccentric orbital longitude instead of the eccentric anomaly. Kepler's equations should be rewritten in terms of the mean and the eccentric orbital longitudes. This fact was recognized by Herget (1962), who used the Laplacian and the Gibbsian vectors as a non-singular set in the computation of special planetary perturbations.

The Hamiltonian vector $\mathbf{h}$ is defined as

$$\mathbf{h}=e\mathbf{Q}/p^{1/2}=u\mathbf{i}+v\mathbf{j} \tag{1}$$

where

$$u=-(e\sin\chi)/p^{1/2} \qquad v=+(e\cos\chi)/p^{1/2} \tag{2}$$

With the introduction the eccentric orbital longitude of

$$\lambda=E+\chi$$

and the unit vector

$$\varrho=\mathbf{i}\cos\lambda+\mathbf{j}\sin\lambda \tag{3}$$

the position vector of the satellite can be written in a form depending directly upon the vector $\mathbf{h}$:

$$\mathbf{r}=a\boldsymbol{\varrho}-\frac{c^2a(a)^{1/2}}{a^{1/2}+c}\mathbf{h}\mathbf{h}\cdot\boldsymbol{\varrho}+ac\mathbf{k}\times\mathbf{h} \tag{4}$$

$$=\xi\mathbf{i}+\eta\mathbf{j}$$

For the determination of p we have

$$1/p=(1/a)+h^2 \tag{5}$$

The semimajor axis remains invariant under secular action.

Eliminating E, M, and e from

$$r/a=1-e\cos E$$

$$E-e\sin E=M$$

in favor of λ, L, and the components of $\mathbf{h}$, we deduce

$$r/a=1-cv\cos\lambda+cu\sin\lambda \tag{6}$$

$$\lambda-cv\sin\lambda-cu\cos\lambda=L \tag{7}$$

In addition to the vectors $\mathbf{r}$ and $\mathbf{h}$, we will also make use of the vector $\mathbf{r}_0$ and $\mathbf{h}_0$ obtained from $\mathbf{r}$ and $\mathbf{h}$ by the substitution of $(\mathbf{i}_0, \mathbf{j}_0, \mathbf{k}_0)$ for $(\mathbf{i}, \mathbf{j}, \mathbf{k})$

$$\mathbf{r}_0=\xi\mathbf{i}_0+\eta\mathbf{j}_0 \qquad \mathbf{h}=u\mathbf{i}_0+v\mathbf{j}_0$$

Decomposing $\mathbf{h}$ along the axes rigidly connected with the osculating orbit plane, we have (Musen, 1954)

$$\frac{d\mathbf{h}}{dt}=\left(1+\frac{r}{p}\right)\mathbf{F}-\frac{\mathbf{r}\mathbf{r}\cdot\mathbf{F}}{rp} \tag{8}$$

Designating by $\mathbf{F}_0$ the disturbing force averaged along the orbit of the disturbing body, we have for the secular effect in $\mathbf{h}$

$$\frac{d\mathbf{h}}{dt}=\frac{1}{2\pi}\int_{-uc}^{2\pi-uc}\left[\left(1+\frac{r}{p}\right)\mathbf{F}_0-\frac{\mathbf{r}\mathbf{r}\cdot\mathbf{F}_0}{rp}\right]dL \tag{9}$$

Eliminating dL in favor of $d\lambda$ by means of (7) we obtain

$$\frac{d\mathbf{h}}{dt}=\frac{1}{2\pi}\int_0^{2\pi}\left[\left(1+\frac{r}{p}\right)\mathbf{F}_0-\frac{\mathbf{r}\mathbf{r}\cdot\mathbf{F}_0}{rp}\right]\frac{r}{a}\,d\lambda \tag{10}$$

We have

$$\mathbf{F}_0=-2m'(\mathbf{i}\mathbf{i}+\mathbf{j}\mathbf{j}\cdot A\cdot\mathbf{r} \tag{11}$$

$$\mathbf{r}\cdot\mathbf{F}_0=-2m'\mathbf{r}\cdot A\cdot\mathbf{r} \tag{12}$$

where A is the fundamental dyadic which was used by Halphen to determine the components of the averaged disturbing force $\mathbf{F}_0$ in the directions of $\mathbf{P}'$, $\mathbf{Q}'$, $\mathbf{R}'$. No attempt is made in this article to modify the method of computing A. Designating the components of $\mathbf{F}_0$ in the directions of $\mathbf{i}$, $\mathbf{j}$, $\mathbf{r}$, by X_0, Y_0, S_0, respectively, we have

$$X_0=-2m'\mathbf{i}\cdot A\cdot\mathbf{r} \tag{13}$$

$$Y_0=-2m'\mathbf{j}\cdot A\cdot\mathbf{r} \tag{14}$$

$$S_0=-2m'\mathbf{r}\cdot A\cdot\mathbf{r}/r \tag{15}$$

and

$$Z_0=-2m'\mathbf{k}\cdot A\cdot\mathbf{r} \tag{16}$$

We deduce from (10), taking (13)–(15) into account, the following two scalar equations for the actual computation of the secular perturbations in $\mathbf{h}$:

$$\frac{du}{dt}=\frac{1}{2\pi}\int_0^{2\pi}\left[\left(1+\frac{r}{p}\right)X_0-\frac{\xi S_0}{p}\right]\frac{r}{a}\,d\lambda \tag{17}$$

$$\frac{dv}{dt}=\frac{1}{2\pi}\int_0^{2\pi}\left[\left(1+\frac{r}{p}\right)Y_0-\frac{\eta S_0}{p}\right]\frac{r}{a}\,d\lambda \tag{18}$$

Introducing the Gibbsian vector $\mathbf{g}$, we can express the vector $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ of the system rigidly connected with the orbit plane in terms of their initial values (i_0, j_0, k_0) and in terms of $\mathbf{g}$. We have (Gibbs, 1901)

$$\mathbf{U}=\mathbf{U}_0+[2/(1+\mathbf{g}^2)]\cdot[\mathbf{g}\times\mathbf{U}_0+\mathbf{g}\times(\mathbf{g}\times\mathbf{U}_0)] \tag{19}$$

where $\mathbf{U}$ and $\mathbf{U}_0$ can take the following values

$$\mathbf{U}=\mathbf{i},\ \mathbf{j},\ \mathbf{k},\ \mathbf{r},\ \mathbf{h}$$

$$\mathbf{U}_0=\mathbf{i}_0,\ \mathbf{j}_0,\ \mathbf{k}_0,\ \mathbf{r}_0,\ \mathbf{h}_0$$

The differential equation for the variation of $\mathbf{g}$ was established in previous work (Musen, 1961). It is

$$d\mathbf{g}/dt=[(1+g^2)/4(p)^{1/2}]Z(\mathbf{r}+\mathbf{r}_0) \tag{20}$$

From this equation we deduce for the secular perturbations in $\mathbf{g}$

$$d\mathbf{g}/dt=\frac{1}{2\pi}\frac{1+g^2}{4(p)^{1/2}}\int_0^{2\pi}Z_0(\mathbf{r}+\mathbf{r}_0)\,(r/a)\,d\lambda \tag{21}$$

Taking into account

$$\mathbf{r}=\mathbf{r}_0+[2/(1+g^2)]\cdot[\mathbf{g}\times\mathbf{r}_0+\mathbf{g}\times(\mathbf{g}\times\mathbf{r}_0)] \tag{22}$$

and, decomposing $\mathbf{g}$ along the axes ($\mathbf{i}_0$, $\mathbf{j}_0$, $\mathbf{k}_0$),

$$\mathbf{g}=g_1\mathbf{i}_0+g_2\mathbf{j}_0+g_3\mathbf{k}_0 \tag{23}$$

we deduce the following three symmetrical scalar equations for the actual computation of the secular effects in $\mathbf{g}$:

$$\frac{dg_1}{dt}=\frac{1}{2\pi}\int_0^{2\pi}\frac{Z_0}{2(p)^{1/2}}[(1+g_1^2)\xi + (g_1g_2-g_3)\eta](r/a)\,d\lambda \tag{24}$$

$$\frac{dg_2}{dt}=\frac{1}{2\pi}\int_0^{2\pi}\frac{Z_0}{2(p)^{1/2}}[(g_1g_2+g_3)\xi + (1+g_2^2)\eta](r/a)\,d\lambda \tag{25}$$

$$\frac{dg_3}{dt}=\frac{1}{2\pi}\int_0^{2\pi}\frac{Z_0}{2(p)^{1/2}}[(g_1g_3-g_2)\xi + (g_2g_3+g_1)\eta](r/a)\,d\lambda \tag{26}$$

The components of the Hamiltonian vector with respect to the inertial system are determined using (19) thus:

$$\mathbf{h}=\mathbf{h}_0+[2/(1+g^2)][\mathbf{g}\times\mathbf{h}_0+\mathbf{g}\times(\mathbf{g}\times\mathbf{h}_0)]$$

The Laplacian vector $\mathbf{e}=e\mathbf{P}$ can be obtained from the relation

$$\mathbf{e}=c\mathbf{h}\times\mathbf{k} \tag{27}$$

If we want to have the standard elements Ω and i, we can easily deduce them from the vectorial equation

$$\mathbf{R}=\mathbf{k}_0+[2/(1+g^2)][\mathbf{g}\times\mathbf{k}_0+\mathbf{g}\times(\mathbf{g}\times\mathbf{k}_0)]$$

For the secular perturbations in the mean orbital longitude we have

$$\frac{dL}{dt}=-\frac{1}{2\pi}\int_0^{2\pi}2\frac{r^2S_0}{a(a)^{1/2}}\,d\lambda+\frac{p(a)^{1/2}}{a^{1/2}+p^{1/2}}\left(u\frac{dv}{dt}-v\frac{du}{dt}\right) \tag{28}$$

In addition to secular effects, the resonance caused by the commensurability of mean motions of the disturbed and the disturbing body is of great significance in the determination of the orbital stability over a long interval of time.

If the process of integration is performed analytically, a small divisor will appear in the denominator as a result of integration of the critical trigonometrical terms.

However, such resonance effects can be determined analytically only under the restriction that the eccentricity and the inclination are small and that only one critical argument is present in the trigonometrical development of the disturbing function. In addition, in the existing analytical theory the secular effect and the resonance effect are treated separately.

In the real problems we might have more than one critical argument, and the mutual effect of the secular and the critical terms might not be negligible.

No restrictions on the eccentricity or the inclination can be imposed in the problem of motion of a satellite in the cislunar space. For all these reasons, I would like to suggest a method of special perturbations for treating the combined secular and resonance effects instead of an analytical development.

Let us assume that

$$n'/n\sim p/p'$$

$$(p,\ p'>0 \quad \text{and are relative primes}) \tag{29}$$

The arguments in the development of the right-hand sides of the equations for the variations of elements are of the form

$$iL-i'M'$$

Using the Liouville (1836) substitution

$$\theta=pL-p'M' \tag{30}$$

$$L=p'\varphi$$

we bring the arguments to the form

$$(ip'-i'p)\varphi+(i'/p')\theta$$

Evidently an argument will be a critical one if

$$i=kp \qquad i'=kp' \qquad (k \text{ is a positive integer}) \tag{31}$$

If the last condition is not satisfied, the term is a short periodic one. This leads to the conclusion that the resonance terms in the equations for variation of $\mathbf{h}$ and $\mathbf{g}$ are obtained by the process of averaging with respect to φ.

Putting

$$\psi^{(u)}(\theta)=\frac{1}{2\pi}\int_0^{2\pi}\left[\left(1+\frac{r}{p}\right)X-\frac{\xi S}{p}\right]d\varphi \tag{32}$$

$$\psi^{(v)}(\theta)=\frac{1}{2\pi}\int_0^{2\pi}\left[\left(1+\frac{r}{p}\right)Y-\frac{\eta S}{p}\right]d\varphi \tag{33}$$

$$\psi^{(g_1)}(\theta)=\frac{1}{2\pi}\int_0^{2\pi}\frac{Z}{2(p)^{1/2}}[(1+g_1^2)\xi + (g_1g_2-g_3)\eta]\,d\varphi \tag{34}$$

$$\psi^{(g_2)}(\theta)=\frac{1}{2\pi}\int_0^{2\pi}\frac{Z}{2(p)^{1/2}}[(g_1g_2+g_3)\xi + (1+g_2^2)\eta]\,d\varphi \tag{35}$$

$$\psi^{(g_3)}(\theta)=\frac{1}{2\pi}\int_0^{2\pi}\frac{Z}{2(p)^{1/2}}[(g_1g_3-g_2)\xi + (g_1+g_3g_2)\eta]\,d\varphi \tag{36}$$

$$\psi^{(n)}(\theta)=-\frac{1}{2\pi}\int_0^{2\pi}\frac{3}{a^{1/2}}\left(\frac{T}{p^{1/2}}+\mathbf{F}\cdot\mathbf{h}\right)d\varphi \tag{37}$$

$$\psi^{(L)}(\theta)=-\frac{1}{2\pi}\int_0\frac{2rS}{a^{1/2}}\,d\varphi+\frac{p(a)^{1/2}}{a^{1/2}+p^{1/2}}[u\psi^{(v)}-v\psi^{(u)}] \tag{38}$$

and

$$A_0{}^{(x)}=\frac{1}{2\pi}\int_0^{2\pi}\psi^{(x)}(\theta)\,d\theta \tag{39}$$

we deduce for the effect of the critical terms a set of equations of the form

$$dx/dt=\psi^{(x)}(\theta)-A_0{}^{(x)}=A_1\cos\theta+B_1\sin\theta + A_2\cos2\theta+B_2\sin2\theta+\cdots \tag{40}$$

In these integrals $d\varphi$ can be replaced by $(1/p')\,(r/a)\,d\lambda$. The upper limit becomes $2p'\pi$.

The computation of the position vector of the disturbing body is then performed using the formulas

$$\mathbf{r}'=\mathbf{P}'a'(\cos E'-e')+\mathbf{Q}'a'(1-e'^2)^{1/2}\sin E' \tag{41}$$

$$E'-e'\sin E'=(p/p')\cdot(\lambda-cv\sin\lambda-cu\cos\lambda) - (1/p')\theta \tag{42}$$

The effect of the critical terms decreases for higher multiples of the argument θ. Assuming that the effect of the critical terms with the arguments 3θ, 4θ, $\cdots$, is negligible, we deduce the following expressions for the coefficients of the first four critical terms

$$A_1=\tfrac{1}{2}[\psi(0)-\psi(\pi)] \tag{43}$$

$$B_1=\tfrac{1}{2}[\psi(\pi/2)-\psi(3\pi/2)] \tag{44}$$

$$A_2=\tfrac{1}{4}[\psi(0)-\psi(\pi/2)+\psi(\pi)-\psi(3\pi/2)] \tag{45}$$

$$B_2=\psi(\pi/4)-\tfrac{1}{4}(\sqrt{2}+1)[\psi(0)+\psi(\pi/2)] + \tfrac{1}{4}(\sqrt{2}-1)[\psi(\pi)+\psi(3\pi/2)] \tag{46}$$

Thus the problem of determining the coefficients of the first four dominant terms is reduced to the computation of four single integrals for each scalar element. There is no necessity to perform a double harmonic analysis in full. This method of reducing the problem to the computation of single integrals was suggested by Liouville (1836), but it was not used before because the numerical evaluation of all these integrals requires a large number of subdivision points and the computation of a large number of special values of X, Y, Z, S, T. Liouville's approach can be used conveniently only now with the advent of modern electronic digital computers.

After the coefficients A_1, B_1, A_2, B_2, $\cdots$, are determined, the equations 40 are combined with the corresponding equations giving the secular derivatives of the elements. The combined equations are then integrated numerically. The value of the critical argument to be substituted into (40) is given by the equation

$$\theta=i[L_0+\Delta L+n_0(t-t_0)] - i'[M_0'+\Delta M_0'+n_0'(t-t_0)] + \int_{t_0}^{t}dt\int_{t_0}^{t}\left(i\frac{dn}{dt}-i'\frac{dn'}{dt}\right)dt \tag{47}$$

CONCLUSION

The combined effect of the secular and the critical terms is decisive in the determination of the orbital stability over a long interval of time.

The combination of Halphen's and Liouville's methods can provide an answer valid at least over an interval of some thousands of years for a minor planet and over a number of years for a

satellite moving in the cislunar space. Our application of the method of secular effects led to some conclusions that were not anticipated beforehand. It remains to be seen what interesting consequences will follow from the superposition of the secular and the critical terms and from the computation of their direct actions as well as their interactions.

BIBLIOGRAPHY

GIBBS, J. W., *Vector Analysis*, pp. 339–347, Yale University Press, New Haven, 1901.

HALPHEN, G. H., *Traité des Fonctions Eliptiques*, vol. 2, Paris, 1888.

HERGET, D., On the variation of arbitrary vectorial constants, *Astron. J.*, **67**(1), 16–18, 1962.

LIOUVILLE, J., Note sur le calcul des inégalités periodiques du mouvement des planètes, *J. Math. Pures Appl.*, p. 197, 1836.

MUSEN, P., Special perturbations of the vectorial elements, *Astron. J.*, **59**(7), 262–267, 1954.

MUSEN, P., On Stromgren's method of special perturbations, *J. Astronaut. Sci.*, **8**(2), 48, 1961.

MUSEN, P., A discussion of Halphen's method, *Rev. Geophys.*, **1**(1), 85–122, 1963.

SHUTE, B. E., A cislunar orbit, *Astron. J.*, **67**, 1962.

STROMGREN, B., *Pub. Med. Kobenhavens Obs.*, **65**, 5, 1929.

A DISCUSSION OF HALPHEN'S METHOD FOR SECULAR PERTURBATIONS AND ITS APPLICATION TO THE DETERMINATION OF LONG RANGE EFFECTS IN THE MOTIONS OF CELESTIAL BODIES. PART I

PETER MUSEN

Goddard Space Flight Center

The long-range effects caused by the moon and the sun are of primary importance in establishing the stability of highly eccentric satellite orbits. At present no complete analytical theory exists which can treat such orbits. It is shown here that Halphen's method of treating secular planetary effects can, by means of step-by-step integration, also be used to determine long range lunar effects in the motions of artificial satellites. Halphen's method permits the numerical integration of long range lunar effects over an interval of a few tens of years. The long range solar effects can be treated by averaging the disturbing function over the orbit of the satellite. Halphen's method is applicable to the determination of long range ("secular") effects in the motion of minor planets over the interval of hundreds of thousands of years. We assume that no sharp commensurability between mean motions of the disturbed and disturbing bodies does exist. A complete theory of Halphen's method is presented in modern symbols. Goursat transformations and a summability process are applied to speed the convergence of series which appear in the theory.

INTRODUCTION

In treating the problem of the orbital stability of celestial bodies, the long-range effects are of primary importance; but no complete analytical theory considering these effects exists at present for large values of the eccentricity, inclination, and semimajor axis. To obtain information about the changes of its orbit over a long time interval, and hence information about the lifetime of, say, an artificial satellite, we have to resort to numerical integration.

Methods based on the use of an unaveraged disturbing function, such as those of Cowell or Encke, contain both the short and long period terms; and in the case of artificial satellites, they require that the interval of integration be much less than the satellite's period, thus creating a large accumulation of round-off errors. The main long range effects in the elements are produced by the long range terms in the disturbing function and by their cross actions. The short period terms can also produce long range effects through their mutual cross actions in higher order approximations, but such effects are very small[1] and can be neglected over a very long time interval. For these reasons, and to diminish the accumulation of round-off errors, it is necessary at the very beginning to remove the short period terms from the disturbing function or from the components of the disturbing force. The problem thus formulated does not differ from the problem of determining the secular perturbation of planets and comets by means of numerical integration using the Gaussian method.[2] With the advent of modern electronic equipment, such a solution of the problem has become possible.

The use of Halphen's form of the Gaussian theory[3] was suggested by the author as a practical method for determining the long-range effects through a step-by-step integration.[4] Previously

Halphen's method was not in use, *probably because of several numerical errors which appear in the original publication. They were all corrected by Goriachev*,[5] whose name should be associated with the method as well; and in its present form the method should properly be called the Halphen-Goriachev method. Some parts of Halphen's original exposition can easily be recognized from the modern standpoint as an application of the calculus of dyadics (matrices) in a hidden form. In the present exposition we shall resort to vectors and matrices. The reason for this is not merely the wish to modernize the notations, but because the application of vectors and matrices removes all the ambiguities and difficulties connected with determining direction cosines when scalars are used. The latter problems are, on some occasions, sources of errors in Halphen's original presentation.

In Goriachev's work, all the formulas given in the final collection are correct; however, there are some misprints in the theoretical exposition. They are corrected here. The author has suggested[4] the use of the Goursat transformation[6] and of the *E*-summability process to speed the convergence of hypergeometric series which appear in the Halphen-Goriachev method and to facilitate the numerical computation.

A group working in celestial mechanics at Goddard Space Flight Center has applied Halphen's method of secular perturbations to the motions of planets, comets, and artificial satellites. The Halphen-Goriachev method was carefully compared with some other existing methods before it was recommended for large scale use. At present, no exposition of Halphen's method exists in English. In undertaking this exposition, the author was also motivated by the wish to present an interesting and important theory to the community of English-speaking astronomers.

SECULAR DISTURBING FUNCTION

The following notation will be used to describe the motion of the disturbed body:

- $\mathbf{r}$ the position vector with respect to the central body,
- $\mathbf{r}^0$ the unit vector in the direction of $\mathbf{r}$,
- $\mathbf{P}$ the unit vector directed from the central body toward the osculating perigee,
- $\mathbf{R}$ the unit vector, normal to the osculating orbit plane, in the direction of angular momentum,
- $\mathbf{Q} = \mathbf{R} \times \mathbf{P}$,
- e the osculating eccentricity,
- a the semimajor axis,
- $b = a(1-e^2)^{1/2}$, the semimajor axis,
- v the true anomaly,
- E the eccentric anomaly,
- g the mean anomaly.

Primed notations will be used to describe the motion of the disturbing body. The position vector of the disturbing body relative to the disturbed body will be designated by $\mathbf{p}$, where $\mathbf{p} = \mathbf{r}' - \mathbf{r}$.

Let m' be the mass of the disturbing body, f be the gravitational constant, and m be the mass of the disturbed body. The mass of the central body will be designated by M.

If the disturbing force

$$\mathbf{F} = fm' \left(\frac{\boldsymbol{\rho}}{\rho^3} - \frac{\mathbf{r}'}{r'^3} \right) \tag{1}$$

is developed into a double Fourier series with arguments g and g', then the constant term in the development is the secular disturbing force $[\mathbf{F}]$ and we have

$$[\mathbf{F}] = \frac{fm'}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} \left(\frac{\boldsymbol{\rho}}{\rho^3} - \frac{\mathbf{r}'}{r'^3} \right) dg\, dg' \,. \tag{2a}$$

Thus $[\mathbf{F}]$ is deduced from the Equation 1 by applying a double process of averaging over the orbit of the disturbing body and over the orbit of the disturbed body.

Writing the "area integral" for the disturbing body in the form

$$dg' = \frac{r'^2\, dv'}{a' b'} \,, \tag{2b}$$

we have

$$\frac{1}{2\pi} \int_0^{2\pi} \frac{\mathbf{r}'}{r'^3}\, dg' = \frac{1}{2\pi a' b'} \int_0^{2\pi} \mathbf{r}^{0\prime}\, dv' = 0 \,.$$

Thus the indirect part $(-\mathbf{r}'/r'^3)$ of the disturbing force does not produce any secular effects and equation 2a takes the form

$$[\mathbf{F}] = \frac{fm'}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} \frac{\boldsymbol{\rho}}{\rho^3}\, dg\, dg' \,. \tag{3}$$

Let us first apply to $\mathbf{F}$ the process of averaging over the orbit of the disturbing body. This

averaged force will be designated by $\mathbf{F}_0$:

$$\mathbf{F}_0 = \frac{fm'}{2\pi}\int_0^{2\pi}\frac{\rho}{\rho^3}\,dg' \cdot \qquad (4)$$

In the process of determining $\mathbf{F}_0$ the position of the disturbing body is imagined to describe the complete osculating ellipse. However, we are interested neither in short period terms nor in knowing at what moment of time the disturbing body will occupy a particular position in its ellipse. This process of averaging is evidently a purely geometrical one.

The geometrical locus of vectors ρ is an elliptical cone with its apex in the disturbed body. Taking equation 2b into account, we can also write

$$\mathbf{F}_0 = \frac{fm'}{2\pi a' b'}\int_0^{2\pi}\frac{\rho}{\rho^3}\,r'^2\,dv' \cdot \qquad (5)$$

If we consider two neighboring position vectors ρ and $\rho+d\rho$ with respect to the disturbed body m, then $r'^2\,dv'/2$ represents the area of the elementary sector with the apex in the central body (fig. 1). Taking equation 5 into account and setting

$$d\mu = \frac{m'\,r'^2\,dv'}{2\pi a' b'} ,$$

we deduce

$$\mathbf{F}_0 = f\oint\frac{\rho}{\rho^3}\,d\mu \cdot \qquad (6)$$

This integral is taken along the ellipse of the disturbing body in the direction of the motion. Equation 6 represents the Gaussian result: $\mathbf{F}_0$ *is equal to the attraction of an elliptic ring over which the mass is distributed proportionally to the area of the sector described by the radius vector* $\mathbf{r}'$.

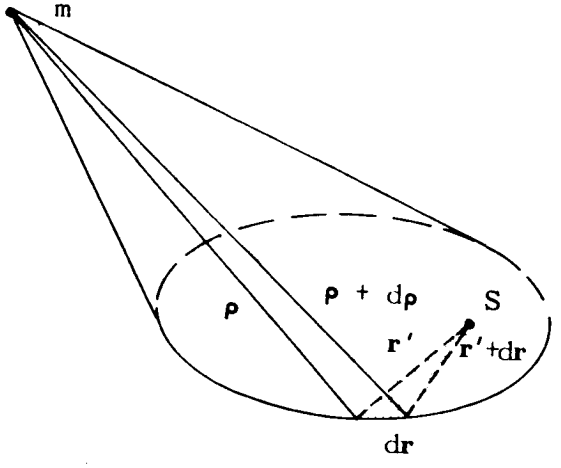


FIGURE 1.—Area of the elementary sector.

Let ρ_0 be the position vector of the central body relative to the disturbed body. Evidently

$$\rho_0 = -\mathbf{r} .$$

Also let

$$h = \rho_0 \cdot \mathbf{R}'$$

be the projection of ρ_0 on $\mathbf{R}'$. We have

$$\rho = \rho_0 + \mathbf{r}' ,$$

$$d\mathbf{r}' = d\rho ,$$

$$\rho_0 \cdot \rho \times d\rho = \rho_0 \cdot (\rho_0 \times d\mathbf{r}' + \mathbf{r}' \times d\mathbf{r}')$$

$$= \rho_0 \cdot \mathbf{R}'\,r'^2\,dv' = hr'^2\,dv'$$

or

$$r'^2\,dv' = \frac{\rho_0 \cdot \rho \times d\rho}{h} .$$

As a result of the last equation, equation 5 becomes

$$\mathbf{F}_0 = \frac{fm'}{2\pi a' b' h}\oint\frac{\rho\,\rho \times d\rho}{\rho^3} \cdot \rho_0 . \qquad (7)$$

Equation 7 introduces a dyadic

$$\Phi = \frac{1}{2}\oint\frac{\mathbf{r}\,\rho \times d\rho}{\rho^3} \qquad (8)$$

closely associated with the problem of determining secular perturbations.

Designating the unit vector in the direction of ρ by ρ^0 and substituting $\rho = \rho\rho^0$ into equation 8, we can reduce Φ to the simpler form

$$\Phi = \frac{1}{2}\oint\rho^0\rho^0 \times d\rho^0 . \qquad (9)$$

The peculiar characteristics of the dyadic Φ are that its first scalar invariant and its vector are both equal to zero.

We have

$$\Phi_s = \frac{1}{2}\oint\rho^0 \cdot \rho^0 \times d\rho^0 = 0 \qquad (10)$$

and

$$\Phi_x = \frac{1}{2}\oint \rho^0 \times (\rho^0 \times d\rho^0) \cdot$$

Developing the double cross-product in the form

$$\Phi_x = \frac{1}{2}\oint (\rho^0 \rho^0 \cdot d\rho^0 - \rho^0 \cdot \rho^0 \, d\rho^0)$$

and taking

$$\rho^0 \cdot d\rho^0 = 0 ,$$

$$\rho^0 \cdot \rho^0 = 1$$

into account, we deduce

$$\Phi_x = -\frac{1}{2}\oint d\rho^0 = 0 . \qquad (11)$$

The condition that the vector of a dyadic equal zero is necessary and sufficient for the symmetry of the dyadic. Consequently, Φ is symmetrical and equation 7 can be written in the form

$$F_0 = \frac{fm' \Phi \cdot \rho_0}{\pi a' b' h} = \frac{fm' \rho_0 \cdot \Phi}{\pi a' b' h} ; \qquad (12)$$

or, taking

$$\operatorname{grad}_{\rho_0} \left(\frac{1}{2} \rho_0 \cdot \Phi \cdot \rho_0\right) = \Phi \cdot \rho_0$$

into account and putting

$$\Psi = \frac{1}{2} \frac{fm'}{\pi a' b' h} \rho_0 \cdot \Phi \cdot \rho_0 , \qquad (13a)$$

we can write

$$F_0 = \operatorname{grad}_{\rho_0} \Psi . \qquad (13b)$$

By substituting

$$\rho = x\,i + y\,j + z\,k ,$$

$$\rho \times d\rho = (y\,dz - z\,dy)\,i + (z\,dx - x\,dz)\,j + (x\,dy - y\,dx)\,k ,$$

into equation 8, the following expressions for the components of Φ are obtained:

$$\Phi_{11} = \frac{1}{2}\oint \frac{x(y\,dz - z\,dy)}{\rho^3} ; \qquad (14)$$

$$\Phi_{22} = \frac{1}{2}\oint \frac{y(z\,dx - x\,dz)}{\rho^3} ; \qquad (15)$$

$$\Phi_{33} = \frac{1}{2}\oint \frac{z(x\,dy - y\,dx)}{\rho^3} ; \qquad (16)$$

$$\Phi_{12} = \Phi_{21} = \frac{1}{2}\oint \frac{x(z\,dx - x\,dz)}{\rho^3} = \frac{1}{2}\oint \frac{y(y\,dz - z\,dy)}{\rho^3} ; \qquad (17)$$

$$\Phi_{23} = \Phi_{32} = \frac{1}{2}\oint \frac{y(x\,dy - y\,dx)}{\rho^3} = \frac{1}{2}\oint \frac{z(z\,dx - x\,dz)}{\rho^3} ; \qquad (18)$$

$$\Phi_{31} = \Phi_{13} = \frac{1}{2}\oint \frac{z(y\,dz - z\,dy)}{\rho^3} = \frac{1}{2}\oint \frac{x(x\,dy - y\,dx)}{\rho^3} ; \qquad (19)$$

$$\rho^2 = x^2 + y^2 + z^2 .$$

The process of integration is performed in the direction of motion of the disturbing body.

We assume that the original system of coordinates and the system attached to the apex of the cone are both right-handed systems. By rotating the system associated with the apex of the cone it is always possible to reduce the equation to the cone to its normal form

$$\frac{x^2}{p} + \frac{y^2}{q} + \frac{z^2}{r} = 0 \qquad (20a)$$

and at the same time to cause all points of the ring to have positive z-coordinates. Also, without loss of generality we can assume that p, q, and r in equation 20a satisfy the conditions $p \leqq q < 0 < r$.

The direction of integration in equations 14–19 is positive in the system defined by the unit vectors $\mathbf{P}'$, $\mathbf{Q}'$, $\mathbf{R}'$. However, it can be positive or negative in the system (i, j, k) of principal axes of the cone. The direction of integration will be positive in both systems if $\mathbf{R}' \cdot \mathbf{k} > 0$ and it will be positive in ($\mathbf{P}'$, $\mathbf{Q}'$, $\mathbf{R}'$) and negative in (i, j, k) if $\mathbf{R}' \cdot \mathbf{k} < 0$.

Let

$$\mathbf{s} = \alpha \mathbf{P}' + \beta \mathbf{Q}' + \gamma \mathbf{R}' \qquad (20b)$$

be the position vector of the apex (of the disturbed body) with respect to the center of the ring. We agreed to choose the direction of $\mathbf{k}$ in such a way that the z-coordinates of points of the portion of the plane limited by the ring will be positive in the system ($\mathbf{i}$ $\mathbf{j}$, $\mathbf{k}$); in other words we must have

$$-\gamma \mathbf{R}' \cdot \mathbf{k} > 0 .$$

From this we conclude that the direction of integration will be positive in both systems if $\gamma < 0$ and it will be positive in the system ($\mathbf{P}'$, $\mathbf{Q}'$, $\mathbf{R}'$) and negative in (i, j, k) if $\gamma > 0$. A simple geometrical drawing will confirm this fact. The direction of integration can always be taken as positive in the system (i, j, k) if the factor

$$-\frac{|\gamma|}{\gamma} = -\operatorname{sign} \gamma$$

is attached to the integral. We shall combine this factor with the factor $fm'/\pi a'b'h$ and we shall postpone its introduction until the development is completed. In order to investigate the form of Φ with respect to the system (i, j, k) let us choose as the contour of integration the ellipse

$$\left.\begin{aligned} \frac{x^2}{p} + \frac{y^2}{q} + \frac{z^2}{r} &= 0 \\ z &= 1 . \end{aligned}\right\} \qquad (21)$$

With the condition imposed by equation 21 the preceding equations 14–19 become

$$\Phi_{11} = -\frac{1}{2}\oint \frac{x\,dy}{\rho^3}, \quad \Phi_{22} = +\frac{1}{2}\oint \frac{y\,dx}{\rho^3}, \quad \Phi_{33} = +\frac{1}{2}\oint \frac{x\,dy - y\,dx}{\rho^3}, \qquad (22)$$

$$\Phi_{12} = +\frac{1}{2}\oint \frac{x\,dx}{\rho^3}, \quad \Phi_{23} = +\frac{1}{2}\oint \frac{dx}{\rho^3}, \quad \Phi_{31} = -\frac{1}{2}\oint \frac{dy}{\rho^3}. \qquad (23)$$

We have, in the different quadrants of the ellipse given by equations 21, the following:

	x	y	dx	dy	− x dy	+ y dx	x dy − y dx	x dx
I	+	+	−	+	−	−	+	−
II	−	+	−	−	−	−	+	+
III	−	−	+	−	−	−	+	−
IV	+	−	+	+	−	−	+	+

From this table and equations 22, we see that

$$\Phi_{11} < 0 , \quad \Phi_{22} < 0 , \quad \Phi_{33} > 0$$

and, because of the symmetry of the contour of integration, each of the integrals in equations 22 is equal to four times the integral taken over the first quarter of the ellipse (equations 21). Combining in equations 23 the elements which are symmetrical with respect to the x and y axes, we also conclude from the above table that

$$\Phi_{12} = \Phi_{23} = \Phi_{31} = 0 ,$$

and Φ takes its normal form

$$\Phi = \Phi_{11} \, \mathbf{i} \, \mathbf{i} + \Phi_{22} \, \mathbf{j} \, \mathbf{j} + \Phi_{33} \, \mathbf{k} \, \mathbf{k}$$

in the system of the principal axes of the cone. (Vectors **i**, **j**, **k** are the unit vectors along the principal axes of the cone.) In other words, both the potential

$$\Psi = \frac{fm'}{\pi a' b' h} \cdot \frac{1}{2} \rho_0 \cdot \Phi \cdot \rho_0 \tag{13a}$$

and the equation of the cone can be reduced to their normal forms simultaneously.

In order to compute Φ_{11}, Φ_{22}, Φ_{33} in the system defined by the principal axes, we can take the curve defined by the equations

$$\frac{x^2}{p} + \frac{y^2}{q} + \frac{z^2}{r} = 0 , \quad x^2 + y^2 + z^2 = 1 \tag{24}$$

as the integration contour. Taking the identities

$$\begin{gathered} (s-p)(q-r) + (s-q)(r-p) + (s-r)(p-q) = 0 , \\ p(s-p)(q-r) + q(s-q)(r-p) + r(s-r)(p-q) = (p-q)(q-r)(r-p) , \end{gathered} \tag{25}$$

into account, we can write the equations of the curve defined by equations 24 in the parametric form:

$$x^2 = \frac{p(s-p)}{(r-p)(p-q)} , \tag{26}$$

$$y^2 = \frac{q(s-q)}{(p-q)(q-r)} , \tag{27}$$

$$z^2 = \frac{r(s-r)}{(q-r)(r-p)} , \tag{28}$$

where s is a variable parameter. Considering the inequalities

$$p < 0 , \quad q < 0 , \quad r > 0 ,$$

$$p - q \leq 0 , \quad q - r < 0 , \quad r - p > 0 ,$$

$$x^2 \geq 0 , \quad y^2 \geq 0 , \quad z^2 \geq 0 ,$$

we conclude from equations 26–28 that s must satisfy the conditions

$$p \leq s \leq q \,, \tag{29}$$

and that the two inequalities

$$pqr > 0 \,, \quad (p-q)\,(q-r)\,(r-p) > 0 \tag{30}$$

are satisfied for p, q, and r.

Taking into account the symmetrical form with respect to the z-axis of the contour of integration and the fact that $\rho = 1$ on this contour, we can write equations 14–16 in the form:

$$\Phi_{11} = 2 \int xyz \,[d(\log z) - d(\log y)] \,, \tag{31}$$

$$\Phi_{22} = 2 \int xyz \,[d(\log x) - d(\log z)] \,, \tag{32}$$

$$\Phi_{33} = 2 \int xyz \,[d(\log y) - d(\log x)] \,. \tag{33}$$

Now the integral is taken along the first quadrant of the contour. On the first quadrant we have $y=0$ for $s=q$ and $x=0$ for $s=p$; and the integration is performed in the positive direction from the point

$$x = + \sqrt{\frac{p}{p-r}} \,, \quad y = 0 \,, \quad z = + \sqrt{\frac{r}{r-p}} \,, \quad \text{or} \quad s = q \,,$$

to the point

$$x = 0 \,, \quad y = + \sqrt{\frac{q}{q-r}} \quad z = + \sqrt{\frac{r}{r-q}} \,, \quad \text{or} \quad s = p \,.$$

We deduce from equations 26–28:

$$d(\log x) = \frac{1}{2} \frac{ds}{s-p} \,, \tag{34}$$

$$d(\log y) = \frac{1}{2} \frac{ds}{s-q} \,, \tag{35}$$

$$d(\log z) = \frac{1}{2} \frac{ds}{s-r} \,. \tag{36}$$

In the first quadrant of the contour, we have

$$xyz = - \frac{\sqrt{pqr}}{(p-q)\,(q-r)\,(r-p)} \sqrt{(s-p)\,(s-q)\,(s-r)} \,. \tag{37}$$

When equations 34–37 are taken into consideration, equations 31–33 become:

$$\Phi_{11} = - \int_q^p \frac{\sqrt{pqr}}{(p-q)\,(r-p)} \sqrt{(s-p)\,(s-q)\,(s-r)} \cdot \frac{ds}{(s-q)\,(s-r)} \tag{38}$$

$$\Phi_{22} = -\int_q^p \frac{\sqrt{p\,q\,r}}{(p-q)(q-r)} \sqrt{(s-p)(s-q)(s-r)} \cdot \frac{ds}{(s-r)(s-p)} \tag{39}$$

$$\Phi_{33} = -\int_q^p \frac{\sqrt{p\,q\,r}}{(q-r)(r-p)} \sqrt{(s-p)(s-q)(s-r)} \cdot \frac{ds}{(s-p)(s-q)} \cdot \tag{40}$$

The sign of the square root is chosen to be negative in order that the conditions $\Phi_{11}<0$, $\Phi_{22}<0$, $\Phi_{33}>0$ be satisfied. As before, the integration is performed in the positive direction along the first quarter of the contour. Putting

$$\left.\begin{aligned} e_1 &= \tfrac{1}{3}(2r-p-q) = r - \tfrac{1}{3}(p+q+r) , \\ e_2 &= \tfrac{1}{3}(2q-r-p) = q - \tfrac{1}{3}(p+q+r) , \\ e_3 &= \tfrac{1}{3}(2p-q-r) = p - \tfrac{1}{3}(p+q+r) , \end{aligned}\right\} \tag{41}$$

we have $e_1+e_2+e_3=0$. From $p \leqq q < r$ and equations 41, we obtain

$$e_1 > e_2 \geq e_3 .$$

Let us now introduce in place of s a new independent variable u, by means of the equation

$$s = p(u) + \tfrac{1}{3}(p+q+r) , \tag{42}$$

where $p(u)$ is the Weierstrass elliptic function satisfying the equation

$$p'^2(u) = 4\big(p(u)-e_1\big)\big(p(u)-e_2\big)\big(p(u)-e_3\big) . \tag{43}$$

Equations 41–43 now become

$$s - p = p(u) - e_3 , \tag{44}$$

$$s - q = p(u) - e_2 , \tag{45}$$

$$s - r = p(u) - e_1 , \tag{46}$$

$$ds = p'(u)\,du , \tag{47}$$

$$(s-p)(s-q)(s-r) = \tfrac{1}{4} p'^2(u) .$$

The process of integration in equations 38–40 is performed in the positive direction from $s=q$ to $s=p$. We have for $s=q$

$$p(u) = e_2 ,$$

and for $s=p$

$$p(u) = e_3 .$$

Consequently in the first quarter of the contour the p-function is decreasing and $p'(u)<0$, or

$$\sqrt{(s-p)(s-q)(s-r)} = -\tfrac{1}{2} p'(u) \tag{48}$$

Substituting equations 44–48 into equations 38–40, we deduce

$$\Phi_{11} = \tfrac{1}{2}\int \frac{\sqrt{p\,q\,r}}{(p-q)(r-p)} \cdot \frac{p'^2(u)\,du}{\big(p(u)-e_2\big)\big(p(u)-e_1\big)} , \tag{49}$$

$$\Phi_{22} = \tfrac{1}{2}\int \frac{\sqrt{p\,q\,r}}{(p-q)(q-r)} \cdot \frac{p'^2(u)\,du}{\big(p(u)-e_3\big)\big(p(u)-e_1\big)} , \tag{50}$$

$$\Phi_{33} = \frac{1}{2}\int \frac{\sqrt{p\,q\,r}}{(q-r)\,(r-p)} \cdot \frac{p'^2(u)\,du}{(p(u)-e_3)\,(p(u)-e_2)}, \tag{51}$$

where the integration is still performed in the positive direction over the first quarter of the contour. Substituting the value of $p'^2(u)$ given by equation 43, we obtain:

$$\frac{1}{2}\Phi_{11} = \int C(q-r)\,[p(u)-e_3]\,du, \tag{52a}$$

$$\frac{1}{2}\Phi_{22} = \int C(r-p)\,[p(u)-e_2]\,du, \tag{53a}$$

$$\frac{1}{2}\Phi_{33} = \int C(p-q)\,[p(u)-e_1]\,du, \tag{54a}$$

where

$$C = \frac{\sqrt{p\,q\,r}}{(p-q)\,(q-r)\,(r-p)},$$

and the path of integration is the same as for equations 49–51. The complete contour is described if the parameter u varies from a certain given value $u=u_0$ to $u=u_0+4\omega$, where 2ω is the real period of $p(u)$:

$$\omega = \int_{e_3}^{e_2} \frac{dx}{\sqrt{4(x-e_1)\,(x-e_2)\,(x-e_3)}}.$$

Taking into account the fact that in our case the integral over the first quarter is equal to one fourth of the integral taken over the whole contour, we have from equations 52a, 53a, and 54a:

$$2\Phi_{11} = \int_{u_0}^{u_0+4\omega} C(q-r)\,[p(u)-e_3]\,du, \tag{52b}$$

$$2\Phi_{22} = \int_{u_0}^{u_0+4\omega} C(r-p)\,[p(u)-e_2]\,du, \tag{53b}$$

$$2\Phi_{33} = \int_{u_0}^{u_0+4\omega} C(p-q)\,[p(u)-e_1]\,du. \tag{54b}$$

Taking into account the relations

$$p(u) = -\zeta'(u),$$

$$\zeta(u+4\omega) - \zeta(u) = 4\eta,$$

$$\eta = \zeta(\omega),$$

between Weierstrass functions, we deduce from equations 52b, 53b, and 54b:

$$\frac{1}{2}\Phi_{11} = -C(q-r)\,(\eta+\omega e_3), \tag{55}$$

$$\frac{1}{2}\Phi_{22} = -C(r-p)\,(\eta+\omega e_2), \tag{56}$$

$$\frac{1}{2}\Phi_{33} = -C(p-q)\,(\eta+\omega e_1). \tag{57}$$

Eliminating e_1, e_2, e_3 from the previous equations by means of equations 41, we obtain:

$$\frac{1}{2}\Phi_{11} = -C(q-r)\left[\eta+\left(p-\frac{p+q+r}{3}\right)\omega\right], \tag{58}$$

$$\frac{1}{2}\Phi_{22} = -C(r-p)\left[\eta+\left(q-\frac{p+q+r}{3}\right)\omega\right], \tag{59}$$

$$\frac{1}{2}\Phi_{33} = -C(p-q)\left[\eta+\left(r-\frac{p+q+r}{3}\right)\omega\right]. \tag{60}$$

Putting

$$\mathbf{M} = C[(q-r)\,\mathbf{i}\mathbf{i} + (r-p)\,\mathbf{j}\mathbf{j} + (p-q)\,\mathbf{k}\mathbf{k}] \tag{61}$$

and

$$\mathbf{N} = C[(q-r)\,p\,\mathbf{i}\mathbf{i} + (r-p)\,q\,\mathbf{j}\mathbf{j} + (p-q)\,r\,\mathbf{k}\mathbf{k}], \tag{62}$$

we have, from equations 58–60,

$$-\frac{1}{2}\Phi = M\eta + N\omega - \frac{p+q+r}{3}M\omega \cdot \qquad (63)$$

The next step will be to express Φ in terms of the fundamental dyadic

$$\Theta = \frac{ii}{p} + \frac{jj}{q} + \frac{kk}{r} \qquad (64)$$

and in terms of its invariants

$$k_1 = p + q + r , \qquad (65)$$

$$k_2 = pq + qr + rp , \qquad (66)$$

$$k_3 = pqr . \qquad (67)$$

The dyadic Θ is closely associated with the cone; its components in the inertial system and its invariants (in terms of these components) can be easily deduced. If the dyadics Φ, M, N are expressed in terms of Θ and its invariants, then the components of Φ, M and N in the inertial system can be obtained with no difficulty. Our final goal in this section will be to obtain an expression for Φ and $\mathbf{F}_0$ in terms of invariants of the cone, and in the system defined by the unit vectors $\mathbf{P}'$, $\mathbf{Q}'$, $\mathbf{R}'$.

The dyadic M can be written in the form of a determinant with dyadical elements:

$$M = C\begin{vmatrix} ii & jj & kk \\ p & q & r \\ 1 & 1 & 1 \end{vmatrix} . \qquad (68)$$

Putting

$$\Lambda = \Theta^{-1} = pii + qjj + rkk ,$$

we deduce from equation 68 that

$$M \cdot \begin{vmatrix} p & q & r \\ 1 & 1 & 1 \\ \frac{1}{p} & \frac{1}{q} & \frac{1}{r} \end{vmatrix} = C\begin{vmatrix} \Lambda & I & \Theta \\ k_1^2 - 2k_2 & k_1 & 3 \\ k_1 & 3 & \frac{k_2}{k_3} \end{vmatrix} , \qquad (69)$$

where I is the idemfactor

$$I = ii + jj + kk .$$

Taking into account that

$$\begin{vmatrix} p & q & r \\ 1 & 1 & 1 \\ \frac{1}{p} & \frac{1}{q} & \frac{1}{r} \end{vmatrix} = -\frac{(p-q)(q-r)(r-p)}{k_3} ,$$

and also that

$$C = \frac{\sqrt{k_3}}{(p-q)(q-r)(r-p)} , \qquad (70)$$

we deduce from equation 69 that

$$M = -\frac{\sqrt{k_3}}{(p-q)^2(q-r)^2(r-p)^2} \cdot \begin{vmatrix} \Lambda & I & \Theta \\ k_1^2 - 2k_2 & k_1 & 3 \\ k_1 k_3 & 3k_3 & k_2 \end{vmatrix} . \qquad (71)$$

However, it follows from equation 41 that

$$e_3 - e_2 = p - q ,$$

$$e_1 - e_3 = r - p ,$$

$$e_2 - e_1 = q - r ,$$

and thus equation 71 becomes

$$M = -\frac{\sqrt{k_3}}{(e_1-e_2)^2\,(e_2-e_3)^2\,(e_3-e_1)^2} \cdot \begin{vmatrix} \Lambda & I & \Theta \\ k_1^2-2k_2 & k_1 & 3 \\ k_1 k_3 & 3k_3 & k_2 \end{vmatrix} . \tag{72}$$

The expression $(e_1-e_2)^2\,(e_2-e_3)^2\,(e_3-e_1)^2$ is the discriminant of the equation

$$4\,(x-e_1)\,(x-e_2)\,(x-e_3) = 4x^3 - g_2 x - g_3 = 0 ,$$

$$-g_2 = 4(e_1 e_2 + e_2 e_3 + e_3 e_1) ,$$

$$g_3 = 4 e_1 e_2 e_3 ,$$

and g_2 and g_3 are expressible as invariants of the dyadic Θ.

We know, from either the theory of elliptic functions or the theory of equations, that

$$(p-q)^2\,(q-r)^2\,(r-p)^2 = (e_1-e_2)^2\,(e_2-e_3)^2\,(e_3-e_1)^2 = \frac{1}{16}\Delta , \tag{73}$$

where

$$\Delta = g_2^3 - 27 g_3^2 . \tag{74}$$

It follows from equation 72, by taking the form of the discriminant (equation 73) into consideration, that

$$\begin{aligned} \frac{\Delta}{16\sqrt{k_3}} M &= (9k_3 - k_1 k_2)\,\Lambda \\ &+ (k_1^2 k_2 - 2k_2^2 - 3k_1 k_3)\, I \\ &+ 2(3k_2 - k_1^2) k_3\, \Theta . \end{aligned} \tag{75}$$

Thus the dyadic M is expressible as a linear combination of dyadics $\Lambda=\Theta^{-1}$, I, and Θ. The left side of equation 75a will be designated by μ in accordance with Halphen's notations:

$$\mu = \frac{\Delta}{16\sqrt{k_3}} M . \tag{75b}$$

The dyadic N (equation 62), like M, can also be written in the form of a determinant with dyadical elements in the first row. From equation 62 we have

$$\frac{N}{C} = \begin{vmatrix} p\,i\,i & q\,j\,j & r\,k\,k \\ p & q & r \\ 1 & 1 & 1 \end{vmatrix} . \tag{76}$$

We deduce from equation 76 that

$$\frac{N}{C}\cdot\begin{vmatrix} 1 & 1 & 1 \\ \frac{1}{p} & \frac{1}{q} & \frac{1}{r} \\ \frac{q-r}{p} & \frac{r-p}{q} & \frac{p-q}{r} \end{vmatrix} = \begin{vmatrix} \Lambda & I & \frac{M}{C} \\ k_1 & 3 & 0 \\ 3 & \frac{k_2}{k_3} & -\frac{(p-q)(q-r)(r-p)}{k_3} \end{vmatrix}; \tag{77}$$

and from equations 73 and 75b we obtain

$$\frac{M}{C} = \frac{16\sqrt{k_3}}{\Delta}\cdot\frac{\mu}{C} = \frac{\mu}{(p-q)(q-r)(r-p)}. \tag{78}$$

The value of the determinant in the left side of equation 77 is

$$\frac{2}{pqr}\cdot\left[(p^2+q^2+r^2)-(pq+qr+rp)\right]$$

$$= \frac{2}{pqr}\cdot\left[(p+q+r)^2-3(pq+qr+rp)\right]$$

$$= \frac{2(k_1^2-3k_2)}{k_3},$$

as can be shown very easily. Substituting the above value into equation 77 and taking equations 73 and 78 into account, we have

$$\frac{2(k_1^2-3k_2)}{k_3}\cdot\frac{N}{C} = \frac{1}{(p-q)(q-r)(r-p)}\begin{vmatrix} \Lambda & I & \mu \\ k_1 & 3 & 0 \\ 3 & \frac{k_2}{k_3} & -\frac{1}{16}\frac{\Delta}{k_3} \end{vmatrix}. \tag{79}$$

By taking $1/C$ in the form (from equations 70 and 73)

$$\frac{1}{C} = \frac{1}{16}\cdot\frac{\Delta}{(p-q)(q-r)(r-p)\sqrt{k_3}},$$

we finally obtain from equation 79

$$\frac{\Delta}{16\sqrt{k_3}}N = \frac{1}{2(k_1^2-3k_2)}\cdot\left[\frac{1}{16}\Lambda(k_1 I-3\Lambda)+(k_1k_2-9k_3)\mu\right]. \tag{80}$$

In proving equations 75a and 80 we followed Halphen's reasoning closely, using dyadics instead of quadratic forms. We have shown that the dyadic N, like M, is a linear combination of three fundamental dyadics (Λ, I, Θ) with the coefficients dependent upon the invariants of Θ.

In our proof we used the system of principal axes of the cone, but the use of dyadics leads immediately to the conclusion that the properties of M and N expressed by means of equations 75a and 80 exist in *any* system of coordinates.

Equations 75a and 80 can be written in the form:

$$M = \frac{16\sqrt{k_3}}{\Delta}\left[(9k_3 - k_1k_2)\,\Lambda + (k_1^2k_2 - 2k_2^2 - 3k_1k_3)\,I + 2k_3\,(3k_2 - k_1^2)\,\Theta\right], \tag{81}$$

$$N = \frac{1}{2(k_1^2 - 3k_2)}\left[(k_1 I - 3\Lambda)\sqrt{k_3} + (k_1k_2 - 9k_3)\,M\right], \tag{82}$$

where, as before,

$$\frac{\Delta}{16} = (e_1 - e_2)^2\,(e_2 - e_3)^2\,(e_3 - e_1)^2$$

$$= (p - q)^2\,(q - r)^2\,(r - p)^2$$

$$= \frac{1}{16}\,(g_2^3 - 27\,g_3^2)\,.$$

Now, the discriminant of any cubic equation of the form

$$x^3 + a_1x^2 + a_2x + a_3 = 0$$

can be written as

$$a_1^2a_2^2 - 4a_1^3a_3 + 18a_1a_2a_3 - 4a_2^3 - 27a_3^2\,.$$

For the equation

$$x^3 - k_1x^2 + k_2x - k_3 = 0\,,$$

whose roots are the reciprocals of the characteristic roots $1/p$, $1/q$, $1/r$, for the cone, we deduce that

$$\frac{\Delta}{16} = k_1^2k_2^2 - 4k_1^3k_3 + 18k_1k_2k_3 - 4k_2^3 - 27k_3^2 \tag{83}$$

In addition, from equations 41 and 65–67, we have

$$-\tfrac{1}{4}\,g_2 = \left(p - \tfrac{1}{3}k_1\right)\left(q - \tfrac{1}{3}k_1\right) + \left(q - \tfrac{1}{3}k_1\right)\left(r - \tfrac{1}{3}k_1\right) + \left(r - \tfrac{1}{3}k_1\right)\left(p - \tfrac{1}{3}k_1\right),$$

$$-\tfrac{1}{4}\,g_3 = \left(\tfrac{1}{3}k_1 - p\right)\left(\tfrac{1}{3}k_1 - q\right)\left(\tfrac{1}{3}k_1 - r\right);$$

and, after some transformations,

$$g_2 = \frac{4}{3}\,(k_1^2 - 3k_2)\,, \tag{84}$$

$$g_3 = \frac{4}{27}\,(2k_1^3 - 9k_1k_2 + 27k_3)\,. \tag{85}$$

The next step will be to obtain an expression for Φ in terms of the invariants of the cone. We deduce from equations 63, 82, 84 and 85 that

$$-\frac{1}{2}\Phi = M\left(\eta - \frac{3}{2}\frac{g_3}{g_2}\omega\right) + \frac{2}{3}\omega\frac{\sqrt{k_3}}{g_2}\left(k_1 I - 3\Lambda\right). \tag{86}$$

A second form of equation 86 that will be useful in further exposition is obtained by eliminating M in favor of μ:

$$\frac{1}{2}\Phi = -\left[\frac{16\mu\sqrt{k_3}}{\Delta}\left(\eta - \frac{3}{2}\frac{g_3}{g_2}\omega\right) + \frac{2}{3}\frac{\sqrt{k_3}}{g_2}\left(k_1 I - 3\Lambda\right)\omega\right]. \tag{87}$$

At this point it will be convenient to attach to Φ the factor

$$-\frac{fm'|\gamma|}{\pi a' b' h \gamma}$$

which was purposely omitted throughout the exposition starting from equation 22. Taking (equation 20b)

$$h = \rho_0 \cdot R' = -s \cdot R' = -(\alpha P' + \beta Q' + \gamma R') \cdot R' = -\gamma$$

into account and defining

$$A = \frac{16}{\pi\Delta}\left(\frac{3}{2}\frac{g_3}{g_2}\omega - \eta\right), \tag{88}$$

$$B = \frac{2\omega}{\pi g_2}, \tag{89}$$

we obtain from equation 87:

$$\Phi = \frac{2fm'\sqrt{k_3}}{a' b'|\gamma|}\left[\mu A - \frac{1}{3}B\left(k_1 I - 3\Lambda\right)\right]. \tag{90}$$

Now we can express the disturbing force F_0, averaged over the orbit of the disturbing body, by the formula

$$F_0 = \Phi \cdot \rho_0$$

or

$$F_0 = -\Phi \cdot r, \tag{91}$$

where Φ is now given by equation 90.

In a following chapter we shall derive expressions for the coefficients A and B in terms of invariants of Θ.

ON THE FORM OF THE BASIC DYADIC Φ IN TERMS OF INVARIANTS OF Θ

Let us choose the system (P', Q', R'), with origin in the center of the ring, as a basic reference system. Let (α, β, γ), as before, be the coordinates of the apex of the cone, (ξ, η, ζ) be the coordinates of a point of the cone and $(x_0, y_0, 0)$ be the coordinates of the intersection of the generating line, passing through (ξ, η, ζ), with the ring. Thus we have

$$\frac{\xi - \alpha}{\alpha - x_0} = \frac{\eta - \beta}{\beta - y_0} = \frac{\zeta - \gamma}{\gamma} \tag{92}$$

and

$$\frac{x_0^2}{a'^2} + \frac{y_0^2}{b'^2} = 1. \tag{93}$$

If the origin of the coordinates is transferred to the apex, then the coordinates of a point of the cone become

$$x = \xi - \alpha, \quad y = \eta - \beta, \quad z = \zeta - \gamma$$

and from equation 92 we have

$$x_0 = \frac{\alpha z - \gamma x}{z} \quad \text{and} \quad y_0 = \frac{\beta z - \gamma y}{z} .$$

Substituting these values into equation 93, we obtain the equation of the cone in the form

$$\frac{z^2}{\gamma^2} - \frac{(\alpha z - \gamma x)^2}{a'^2 \gamma^2} - \frac{(\beta z - \gamma y)^2}{b'^2 \gamma^2} = 0 . \quad (94)$$

The divisor γ^2 is introduced for reasons of homogeneity and the condition

$$p \leq q < r$$

requires an arrangement of signs as in equation 94. Therefore this equation can also be written in the form:

$$\mathbf{w} \cdot \Theta \cdot \mathbf{w} = 0 ,$$

where

$$\mathbf{w} = x\mathbf{P}' + y\mathbf{Q}' + z\mathbf{R}' \quad (95)$$

and

$$\Theta = \frac{\mathbf{R}'\mathbf{R}'}{\gamma^2} - \frac{(\alpha\mathbf{R}' - \gamma\mathbf{P}')(\alpha\mathbf{R}' - \gamma\mathbf{P}')}{a'^2\gamma^2} - \frac{(\beta\mathbf{R}' - \gamma\mathbf{Q}')(\beta\mathbf{R}' - \gamma\mathbf{Q}')}{b'^2\gamma^2} . \quad (96)$$

By again introducing $\mathbf{s}$, the position vector of the apex with respect to the center of the ring

$$\mathbf{s} = \alpha\mathbf{P}' + \beta\mathbf{Q}' + \gamma\mathbf{R}' , \quad (20b)$$

we can write the dyadic Θ (equation 96) in a contracted form:

$$\Theta = \frac{\mathbf{s} \times \mathbf{Q}'\mathbf{Q}' \times \mathbf{s}}{a'^2\gamma^2} + \frac{\mathbf{s} \times \mathbf{P}'\mathbf{P}' \times \mathbf{s}}{b'^2\gamma^2} + \frac{\mathbf{R}'\mathbf{R}'}{\gamma^2} , \quad (97)$$

which is more convenient for the computation of invariants.

The form of Θ which we have used previously was

$$\Theta = \frac{\mathbf{i}\,\mathbf{i}}{p} + \frac{\mathbf{j}\,\mathbf{j}}{q} + \frac{\mathbf{k}\,\mathbf{k}}{r} \quad (64)$$

and it referred to the principal axis of the cone.

In the process of computing the invariants we will follow the classical Gibbs notations[7] closely. If a dyadic X is given in a reduced form,

$$X = a\mathbf{l} + b\mathbf{m} + c\mathbf{n} , \quad (98)$$

then the adjoined dyadic X_2 and the invariants X_s, $(X_2)_s$, X_3 are given by the formulas[7]:

$$X_2 = b \times c\mathbf{m} \times \mathbf{n} + c \times a\mathbf{n} \times \mathbf{l} + a \times b\mathbf{l} \times \mathbf{m} , \quad (99)$$

$$X_s = a \cdot \mathbf{l} + b \cdot \mathbf{m} + c \cdot \mathbf{n} , \quad (100)$$

$$(X_2)_s = b \times c \cdot \mathbf{m} \times \mathbf{n} + c \times a \cdot \mathbf{n} \times \mathbf{l} + a \times b \cdot \mathbf{l} \times \mathbf{m} , \quad (101)$$

$$X_3 = (a \cdot b \times c)(\mathbf{l} \cdot \mathbf{m} \times \mathbf{n}) . \quad (102)$$

The reciprocal dyadic is given by the formula:

$$X^{-1} = \frac{X_2}{X_3} . \tag{103}$$

Substituting into equations 99–103 first

$$\mathbf{a} = \frac{\mathbf{s} \times \mathbf{Q}'}{a'^2 \gamma^2} , \quad \mathbf{b} = \frac{\mathbf{s} \times \mathbf{P}'}{b'^2 \gamma^2} , \quad \mathbf{c} = \frac{\mathbf{R}'}{\gamma^2} ,$$

$$\mathbf{l} = \mathbf{Q}' \times \mathbf{s} , \quad \mathbf{m} = \mathbf{P}' \times \mathbf{s} , \quad \mathbf{n} = \mathbf{R}' ,$$

and then

$$\mathbf{a} = \frac{\mathbf{i}}{p} , \quad \mathbf{b} = \frac{\mathbf{j}}{q} , \quad \mathbf{c} = \frac{\mathbf{k}}{r} ,$$

$$\mathbf{l} = \mathbf{i} , \quad \mathbf{m} = \mathbf{j} , \quad \mathbf{n} = \mathbf{k} ,$$

we compute Θ_2, Θ_s, $(\Theta_2)_s$, Θ_3, and Θ^{-1} using both forms of Θ as given by equations 97 and 64, and compare the results. After the substitution and some easy vectorial transformations based on

$$\left.\begin{aligned} \mathbf{a} \times (\mathbf{b} \times \mathbf{c}) &= \mathbf{b}\mathbf{a} \cdot \mathbf{c} - \mathbf{c}\mathbf{a} \cdot \mathbf{b} \\ (\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) &= \mathbf{a} \cdot \mathbf{c}\,\mathbf{b} \cdot \mathbf{d} - \mathbf{a} \cdot \mathbf{d}\,\mathbf{b} \cdot \mathbf{c} \end{aligned}\right\} \tag{104}$$

are performed, we obtain:

$$\Theta_2 = -\frac{\mathbf{P}'\mathbf{P}'}{b'^2 \gamma^2} - \frac{\mathbf{Q}'\mathbf{Q}'}{a'^2 \gamma^2} + \frac{\mathbf{s}\,\mathbf{s}}{a'^2 b'^2 \gamma^2} = \frac{\mathbf{i}\,\mathbf{i}}{q\,r} + \frac{\mathbf{j}\,\mathbf{j}}{r\,p} + \frac{\mathbf{k}\,\mathbf{k}}{p\,q} , \tag{105}$$

$$\Theta_s = \frac{1}{a'^2 b'^2 \gamma^2} \left[a'^2 b'^2 - \alpha^2 b'^2 - \beta^2 a'^2 - \gamma^2 \left(a'^2 + b'^2 \right) \right] = \frac{k_2}{k_3} , \tag{106}$$

$$(\Theta_2)_s = -\frac{1}{b'^2 \gamma^2} - \frac{1}{a'^2 \gamma^2} + \frac{\alpha^2 + \beta^2 + \gamma^2}{a'^2 b'^2 \gamma^2} = \frac{k_1}{k_3} , \tag{107}$$

$$\Theta_3 = \frac{1}{a'^2 b'^2 \gamma^2} = \frac{1}{k_3} , \tag{108}$$

$$\Lambda = \Theta^{-1} = \mathbf{s}\,\mathbf{s} - a'^2\,\mathbf{P}'\mathbf{P}' - b'^2\,\mathbf{Q}'\mathbf{Q}' . \tag{109}$$

From equations 106–108 we deduce expressions for the invariants k_1, k_2, k_3:

$$k_1 = \alpha^2 + \beta^2 + \gamma^2 - \left(a'^2 + b'^2 \right) , \tag{110}$$

$$k_2 = a'^2 b'^2 - \alpha^2 b'^2 - \beta^2 a'^2 - \gamma^2 \left(a'^2 + b'^2 \right) , \tag{111}$$

$$k_3 = a'^2 b'^2 \gamma^2 . \tag{112}$$

The value $(k_3)^{1/2}$ in equation 90 is positive. Consequently, from equation 112,

$$\sqrt{k_3} = a' b' |\gamma| .$$

Substituting this value into equation 90, we obtain a final form of Φ:

$$\Phi = 2 f m' (A\mu + B\nu) , \tag{113}$$

where

$$\nu = \Lambda - \frac{1}{3} k_1 I . \tag{114a}$$

The expression (equation 91) for the disturbing force averaged over the orbit of the disturbing body now becomes

$$\mathbf{F}_0 = - 2 f m' (A\mu \cdot \mathbf{r} + B\nu \cdot \mathbf{r}) . \tag{114b}$$

Using equation 96 we can write Θ in the form

$$\begin{aligned} k_3 \Theta = & - b'^2 \gamma^2 \mathbf{P}' \mathbf{P}' + b'^2 \alpha\gamma \mathbf{P}' \mathbf{R}' \\ & - a'^2 \gamma^2 \mathbf{Q}' \mathbf{Q}' + a'^2 \beta\gamma \mathbf{Q}' \mathbf{R}' \\ & + b'^2 \alpha\gamma \mathbf{R}' \mathbf{P}' + a'^2 \gamma\beta \mathbf{R}' \mathbf{Q}' + \left(a'^2 b'^2 - \alpha^2 b'^2 - \beta^2 a'^2\right) \mathbf{R}' \mathbf{R}' . \end{aligned} \tag{115}$$

In a similar way, from equation 109 we obtain

$$\begin{aligned} \Lambda = & \left(\alpha^2 - a'^2\right) \mathbf{P}' \mathbf{P}' + \alpha\beta \mathbf{P}' \mathbf{Q}' + \alpha\gamma \mathbf{P}' \mathbf{R}' \\ & + \alpha\beta \mathbf{Q}' \mathbf{P}' + \left(\beta^2 - b'^2\right) \mathbf{Q}' \mathbf{Q}' + \beta\gamma \mathbf{Q}' \mathbf{R}' \\ & + \alpha\gamma \mathbf{R}' \mathbf{P}' + \gamma\beta \mathbf{R}' \mathbf{Q}' + \gamma^2 \mathbf{R}' \mathbf{R}' ; \end{aligned} \tag{116}$$

and in addition we have

$$I = \mathbf{P}' \mathbf{P}' + \mathbf{Q}' \mathbf{Q}' + \mathbf{R}' \mathbf{R}' . \tag{117}$$

Now, substituting the results given by equations 115–117 into 75a and 114a, we obtain μ and ν in the form of Halphen's matrices

$$\mu = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \\ \mu_{31} & \mu_{32} & \mu_{33} \end{bmatrix} , \tag{118}$$

$$\nu = \begin{bmatrix} \nu_{11} & \nu_{12} & \nu_{13} \\ \nu_{21} & \nu_{22} & \nu_{23} \\ \nu_{31} & \nu_{32} & \nu_{33} \end{bmatrix}, \tag{119}$$

where

$$\begin{aligned}
\mu_{11} &= (9k_3 - k_1 k_2)(\alpha^2 - a'^2) + k_1(k_1 k_2 - 3k_3) - 2k_2^2 + \frac{3}{2} g_2 b'^2 \gamma^2 , \\
\mu_{22} &= (9k_3 - k_1 k_2)(\beta^2 - b'^2) + k_1(k_1 k_2 - 3k_3) - 2k_2^2 + \frac{3}{2} g_2 a'^2 \gamma^2 , \\
\mu_{33} &= (9k_3 - k_1 k_2)\gamma^2 + k_1(k_1 k_2 - 3k_3) - 2k_2^2 + \frac{3}{2} g_2 (\alpha^2 b'^2 + \beta^2 a'^2 - a'^2 b'^2) , \\
\mu_{12} &= \mu_{21} = (9k_3 - k_1 k_2)\alpha\beta , \\
\mu_{23} &= \mu_{32} = (9k_3 - k_1 k_2)\beta\gamma - \frac{3}{2} g_2 a'^2 \beta\gamma , \\
\mu_{31} &= \mu_{13} = (9k_3 - k_1 k_2)\gamma\alpha - \frac{3}{2} g_2 b'^2 \gamma\alpha ;
\end{aligned} \tag{120}$$

and

$$\begin{aligned}
\nu_{11} &= \alpha^2 - a'^2 - \frac{1}{3} k_1 , \\
\nu_{22} &= \beta^2 - b'^2 - \frac{1}{3} k_1 , \\
\nu_{33} &= \gamma^2 - \frac{1}{3} k_1 , \\
\nu_{12} &= \nu_{21} = \alpha\beta , \\
\nu_{23} &= \nu_{32} = \beta\gamma , \\
\nu_{31} &= \nu_{13} = \gamma\alpha .
\end{aligned} \tag{121}$$

The decomposition of Φ in the system $(\mathbf{P}', \mathbf{Q}', \mathbf{R}')$ takes the form

$$\Phi_{ij} = 2 f m' (A\mu_{ij} + B\nu_{ij}). \tag{122}$$

Putting

$$x = \mathbf{r} \cdot \mathbf{P}' , \quad y = \mathbf{r} \cdot \mathbf{Q}' , \quad z = \mathbf{r} \cdot \mathbf{R}' , \tag{123}$$

we obtain a decomposition of $-\mathbf{F}_0$:

$$-F_{01} = \Phi_{11}x + \Phi_{12}y + \Phi_{13}z\,, \tag{124}$$

$$-F_{02} = \Phi_{21}x + \Phi_{22}y + \Phi_{23}z\,, \tag{125}$$

$$-F_{03} = \Phi_{31}x + \Phi_{32}y + \Phi_{33}z\,. \tag{126}$$

The system of equations 124–126 gives Halphen's decomposition of $\mathbf{F}_0$ along the axes ($\mathbf{P}'$, $\mathbf{Q}'$, $\mathbf{R}'$). Here, however, the author would like to suggest a slightly different system of formulas, which might be simpler to program for computers. This system is based on decompositions of Θ and Λ as given by equations 97 and 109. Putting

$$\begin{aligned}(\Theta_1) &= k_3\mathbf{P}'\cdot\Theta\cdot\mathbf{r}\,,\\ (\Theta_2) &= k_3\mathbf{Q}'\cdot\Theta\cdot\mathbf{r}\,,\\ (\Theta_3) &= k_3\mathbf{R}'\cdot\Theta\cdot\mathbf{r}\,,\end{aligned} \tag{127}$$

$$\begin{aligned}(\Lambda_1) &= \mathbf{P}'\cdot\Lambda\cdot\mathbf{r}\,,\\ (\Lambda_2) &= \mathbf{Q}'\cdot\Lambda\cdot\mathbf{r}\,,\\ (\Lambda_3) &= \mathbf{R}'\cdot\Lambda\cdot\mathbf{r}\,,\end{aligned} \tag{128}$$

$$\begin{aligned}x &= \mathbf{P}'\cdot\mathbf{r}\,,\\ y &= \mathbf{Q}'\cdot\mathbf{r}\,,\\ z &= \mathbf{R}'\cdot\mathbf{r}\,,\end{aligned}$$

and taking the equations

$$\begin{aligned}\mathbf{s} &= \mathbf{r} + a'e'\mathbf{P}'\,,\\ \mathbf{s}\times\mathbf{r} &= a'e'\mathbf{P}'\times\mathbf{r}\end{aligned} \tag{129}$$

into account, we deduce from equations 97 and 109:

$$\begin{aligned}(\Theta_1) &= a'b'^2e'\gamma^2\,,\\ (\Theta_2) &= 0\,,\\ (\Theta_3) &= a'b'^2\gamma(a'-\alpha e')\,,\end{aligned} \tag{130}$$

$$\begin{aligned}(\Lambda_1) &= \alpha\mathbf{s}\cdot\mathbf{r} - a'^2x\,,\\ (\Lambda_2) &= \beta\mathbf{s}\cdot\mathbf{r} - b'^2y\,,\\ (\Lambda_3) &= \gamma\mathbf{s}\cdot\mathbf{r}\,.\end{aligned} \tag{131}$$

Now let us introduce two vectors $\mathbf{m}$ and $\mathbf{n}$ by means of the formulas:

$$\mathbf{m} = \begin{bmatrix}(\Lambda_1) & x & (\Theta_1)\\ (\Lambda_2) & y & (\Theta_2)\\ (\Lambda_3) & z & (\Theta_3)\end{bmatrix}\cdot\begin{bmatrix}9k_3 - k_1k_2\\ k_1(k_1k_2 - 3k_3) - 2k_2^2\\ -\frac{3}{2}g_2\end{bmatrix} = \begin{bmatrix}m_1\\ m_2\\ m_3\end{bmatrix}, \tag{132}$$

$$\mathbf{n} = \begin{bmatrix}(\Lambda_1) - \frac{1}{3}k_1x\\ (\Lambda_2) - \frac{1}{3}k_1y\\ (\Lambda_3) - \frac{1}{3}k_1z\end{bmatrix} = \begin{bmatrix}n_1\\ n_2\\ n_3\end{bmatrix}. \tag{133}$$

By taking equations 75a, 75b and 114a into account, the expressions for the components F_{01}, F_{02}, F_{03} take a more concise form:

$$\left.\begin{aligned} F_{01} &= -2fm'(Am_1 + Bn_1), \\ F_{02} &= -2fm'(Am_2 + Bn_2), \\ F_{03} &= -2fm'(Am_3 + Bn_3). \end{aligned}\right\} \tag{134}$$

EXPRESSIONS FOR THE COEFFICIENTS A AND B IN TERMS OF INVARIANTS OF Θ

In this chapter we present Halphen's theory concerning coefficients A and B with some modifications which speed up convergence of the series. Our final goal will be to express the coefficients A and B (as given by equations 88 and 89) in terms of invariants k_1, k_2, k_3 of Θ. In order to do so we have to make use of some theorems from the theory of the Weierstrass elliptic functions. We have the basic relations

$$p'^2(u) = 4p^3(u) - g_2 p(u) - g_3, \tag{135a}$$

$$p''(u) = 6p^2(u) - \frac{1}{2} g_2; \tag{135b}$$

and we also make use of the equation

$$g_2^2 \frac{\partial \log \sigma(u)}{\partial g_3} + 18 g_3 \frac{\partial \log \sigma(u)}{\partial g_2} = \frac{3}{2} \frac{1}{\sigma(u)} \frac{\partial^2 \sigma(u)}{\partial u^2} + \frac{1}{8} g_2 u^2 \tag{136}$$

(reference 8, p. 393).

From the defining equation for the ζ-function,

$$\frac{1}{\sigma(u)} \frac{\partial \sigma(u)}{\partial u} = \zeta(u), \tag{137}$$

we deduce that

$$\frac{1}{\sigma(u)} \frac{\partial^2 \sigma(u)}{\partial u^2} - \frac{1}{\sigma^2(u)} \left(\frac{\partial \sigma(u)}{\partial u}\right)^2 = \zeta'(u); \tag{138}$$

but

$$\zeta'(u) = -p(u), \tag{139}$$

and equation 138 can be written in the form:

$$\frac{1}{\sigma(u)} \frac{\partial^2 \sigma(u)}{\partial u^2} = \zeta^2(u) - p(u). \tag{140}$$

Substituting this result into equation 136 we obtain

$$g_2^2 \frac{\partial \log \sigma(u)}{\partial g_3} + 18 g_3 \frac{\partial \log \sigma(u)}{\partial g_2} = \frac{3}{2} \zeta^2(u) - \frac{3}{2} p(u) + \frac{1}{8} g_2 u^2. \tag{141}$$

Introducing Halphen's linear differential operator D, defined by the equation

$$D = 12 g_3 \frac{\partial}{\partial g_2} + \frac{2}{3} g_2^2 \frac{\partial}{\partial g_3}, \tag{142}$$

we can write equation 141 in the form:

$$D \log \sigma(u) = \zeta^2(u) - p(u) + \frac{1}{12} g_2 u^2 . \tag{143}$$

Differentiating this result twice with respect to u and taking the defining equations 137 and 139 into account, we deduce that

$$D\zeta(u) = -2\zeta(u)\, p(u) - p'(u) + \frac{1}{6} g_2 u \tag{144}$$

and

$$Dp(u) = 2p'(u)\, \zeta(u) - 2p^2(u) + p''(u) - \frac{1}{6} g_2 . \tag{145}$$

Taking equation 135b into account, we can reduce the last result to

$$Dp(u) = 2p'(u)\, \zeta(u) + 4p^2(u) - \frac{2}{3} g_2 . \tag{146}$$

Differentiating equation 146 with respect to u we obtain

$$Dp'(u) = 6p(u)\, p'(u) + 2p''(u)\, \zeta(u). \tag{147}$$

If we have a function F of g_2, g_3 of the form

$$w = F(v, g_2, g_3) ,$$

and if $v = \phi(g_2, g_3)$ is also a function of g_2, g_3, then

$$Dw = \frac{\partial F}{\partial v} Dv + DF . \tag{148}$$

In forming DF on the right side of the last equation we consider v as a constant.

Designating, as before, the half of the real period by ω, we have

$$p'(\omega, g_2, g_3) = 0 , \quad \eta = \zeta(\omega, g_2, g_3) . \tag{149}$$

Putting $v = \omega$, $w = 0$, $F = p'(\omega, g_2, g_3)$ into equation 148, we obtain

$$p''(\omega, g_2, g_3)\, D\omega + Dp'(\omega) = 0 ;$$

and from the last equation it follows, by taking equations 147 and 149 into account, that

$$p''(\omega, g_2, g_3)\, D\omega + 2\eta p''(\omega, g_2, g_3) = 0$$

or

$$D\omega = -2\eta . \tag{150a}$$

Setting $w = \eta$, $F = \zeta(v, g_2, g_3)$, $v = \omega$ in equation 148, we have

$$D\eta = \zeta'(\omega, g_2, g_3)\, D\omega + D\zeta(\omega, g_2, g_3) . \tag{150b}$$

From equation 144 it follows, by taking equation 149 into account, that

$$D\zeta(\omega, g_2, g_3) = -2\eta p(\omega, g_2, g_3) + \frac{1}{6} g_2 \omega .$$

Substituting this into equation 150b and taking equations 139 and 150a into consideration we obtain

$$D\eta = +2\eta p(\omega, g_2, g_3) - 2\eta p(\omega, g_2, g_3) + \frac{1}{6} g_2 \omega ;$$

or, finally

$$D\eta = +\frac{1}{6} g_2 \omega . \qquad (151)$$

Writing the formula

$$p\left(\frac{u}{\sqrt{\mu}}, \mu^2 g_2, \mu^3 g_3\right) = \mu p(u, g_2, g_3)$$

in the form

$$p(\nu u, \nu^{-4} g_2, \nu^{-6} g_3) = \nu^{-2} p(u, g_2, g_3) ,$$

we conclude that if the argument u is considered to be of the first dimension, then g_2 is of the minus fourth dimension and g_3 is of the minus sixth dimension. The discriminant $\Delta = g_2^3 - 27 g_3^2$ is of the minus twelfth dimension and the absolute invariant $J = g_2^3/\Delta$ is of zero dimension. The real semi-period ω, being an argument, can be considered of the first dimension and, consequently, $x = \omega \Delta^{1/12}$ is of zero dimension. The function $p(u, g_2, g_3)$ is of the minus second dimension and then $\zeta(u, g_2, g_3)$, because of the equation $\zeta'(u) = -p(u)$, must be of the minus first dimension; in particular, $\eta = \zeta(u)$ is also of the minus first dimension and, consequently, $y' = \eta \Delta^{-1/12}$ is of zero dimension. As a consequence x and y' can be treated as functions of only the absolute invariant J. By applying the operator D to Δ and J, we deduce that

$$D\Delta = 0 ,$$
$$DJ = \frac{36 g_2^2 g_3}{\Delta} . \qquad (152)$$

Eliminating g_2 and g_3 by means of the equations $J = g_2^3/\Delta$ and $J - 1 = 27 g_3^2/\Delta$, from equation 152, we obtain

$$DJ = 4\sqrt{3}\, \Delta^{1/6} J^{2/3} (J-1)^{1/2} . \qquad (153)$$

We have, taking equations 152 and 153 into account:

$$Dx = \frac{dx}{dJ} DJ = 4\sqrt{3} \frac{dx}{dJ} \Delta^{1/6} J^{2/3} (J-1)^{1/2} ; \qquad (154)$$

$$Dx = \Delta^{1/12} D\omega . \qquad (155)$$

By substituting $D\omega = -2\eta = -2\Delta^{1/12} y'$ (equation 150a and definition of y') into equation 155 and comparing the result thus obtained with the result given by equation 154, we have

$$\frac{dx}{dJ} = -\frac{1}{2\sqrt{3}} (J-1)^{-1/2} J^{-2/3} y' . \qquad (156)$$

Taking equation 151 into account we deduce that

$$Dy' = D\left(\eta \Delta^{-1/12}\right) = \Delta^{-1/12} D(\eta) = \frac{1}{6} g_2 \omega \Delta^{-1/12} . \qquad (157)$$

Similarly to equation 154, we have for y'

$$Dy' = \frac{dy'}{dJ} DJ = 4\sqrt{3} \frac{dy'}{dJ} \Delta^{1/6} J^{2/3} (J-1)^{1/2} . \qquad (158)$$

Comparing the two values of Dy', as given by equations 157 and 158, we obtain

$$\frac{dy'}{dJ} = \frac{1}{24\sqrt{3}} \Delta^{-1/4} g_2 \omega J^{-2/3} (J-1)^{-1/2} . \qquad (159)$$

Eliminating g_2 and ω from the last equation in favor of J, Δ and x by means of the relations $g_2 = \Delta^{1/3} J^{1/3}$, $\omega = x \Delta^{-1/12}$, we obtain

$$\frac{dy'}{dJ} = \frac{1}{24\sqrt{3}} (J-1)^{-1/2} J^{-1/3} x . \qquad (160)$$

If we put

$$\frac{1}{2\sqrt{3}} (J-1)^{-1/2} J^{-2/3} y' = y , \qquad (161)$$

then equation 156 takes the form

$$\frac{dx}{dJ} = -y . \tag{162}$$

Eliminating y' from equation 160 by means of the defining equation 161, we deduce that

$$144\,J(J-1)\frac{dy}{dJ} + 24\,y(7J-4) - x = 0 . \tag{163}$$

Eliminating y by means of equation 162, we now have

$$J(1-J)\frac{d^2x}{dJ^2} + \left(\frac{2}{3} - \frac{7}{6}J\right)\frac{dx}{dJ} - \frac{1}{144}x = 0 . \tag{164}$$

The last equation has the form of the hypergeometric equation

$$x(1-x)\,y'' + \left[\gamma - (\alpha+\beta+1)\,x\right]y' - \alpha\beta y = 0$$

with $\alpha=\beta=\frac{1}{12}$ and $\gamma=\frac{2}{3}$. Differentiating equation 163 and again taking equation 162 into account, we obtain a hypergeometric equation for y:

$$J(1-J)\frac{d^2y}{dJ^2} + \left(\frac{5}{3} - \frac{19}{6}J\right)\frac{dy}{dJ} - \frac{169}{144}y = 0 \tag{165}$$

with $\alpha=\beta=\frac{13}{12}$ and $\gamma=\frac{5}{3}$.

For the purpose of determining the coefficients A and B (as given by equations 88 and 89) it will be convenient to use

$$X = \omega g_2^{1/4} \tag{166}$$

and

$$Y = \eta g_2^{-1/4} \tag{167}$$

rather than x and y. The next problem will be to determine hypergeometric equations of which X and Y are solutions.

Eliminating ω, η, g_2 from equations 166 and 167 in favor of x, y, J by using

$$x = \omega\Delta^{1/12}, \quad y' = \eta\Delta^{-1/12}, \quad g_2 = J^{1/3}\Delta^{1/3}$$

and equation 161, we obtain

$$x = J^{-1/12}X , \tag{168}$$

$$y = \frac{1}{2\sqrt{3}}(J-1)^{-1/2}J^{-7/12}Y . \tag{169}$$

Also, we introduce

$$\xi = \frac{J-1}{J} \tag{170}$$

as a new independent variable instead of J.

The transformation of equation 164 by the introduction of X instead of x and of ξ instead of J is a transformation of the form $y=x^{-\alpha}z$, $u=(x-1)/x$ applied to the hypergeometric equation

$$x(1-x)\,y'' + \left[\gamma - (\alpha+\beta+1)\,x\right]y' - \alpha\beta y = 0 .$$

We have $x=1/(1-u)$; and by substituting

$$y = (1-u)^{\alpha}z ,$$

$$\frac{dy}{dx} = (1-u)^{\alpha+2}\frac{dz}{du} - \alpha(1-u)^{\alpha+1}z ,$$

$$\frac{d^2y}{dx^2} = (1-u)^{\alpha+4}\frac{d^2z}{du^2} - (2\alpha+2)(1-u)^{\alpha+3}\frac{dz}{du} + (\alpha^2+\alpha)(1-u)^{\alpha+2}z ,$$

into the original equation, we deduce a hypergeometric equation

$$u(1-u)\frac{d^2 z}{du^2} + \left[\gamma_1 - (\alpha_1 + \beta_1 + 1)\, u\right]\frac{dz}{du} - \alpha_1 \beta_1 z = 0$$

with

$$\alpha_1 = \alpha ,$$

$$\beta_1 = \alpha - \gamma + 1 ,$$

$$\gamma_1 = \alpha + \beta - \gamma + 1 .$$

In our case $\alpha=\beta=\frac{1}{12}$, $\gamma=\frac{2}{3}$, and $\alpha_1=\frac{1}{12}$, $\beta_1=\frac{5}{12}$, $\gamma_1=\frac{1}{2}$, and the hypergeometric equation of which X is an integral is

$$\xi(1-\xi)\frac{d^2 X}{d\xi^2} + \left(\frac{1}{2} - \frac{3}{2}\xi\right)\frac{dX}{d\xi} - \frac{5}{144} X = 0 . \tag{171}$$

A hypergeometric equation

$$x(1-x)\, y'' + \left[\gamma - (\alpha+\beta+1)\, x\right] y' - \alpha\beta y = 0$$

has the integrals

$$y_1 = F(\alpha, \beta, \gamma, x) , \tag{172}$$

$$y_2 = x^{1-\gamma} F(\alpha-\gamma+1, \beta-\gamma+1, 2-\gamma, x) , \tag{173}$$

$$y_3 = F(\alpha, \beta, \alpha+\beta-\gamma+1, 1-x) , \tag{174}$$

$$y_4 = (1-x)^{\gamma-\alpha-\beta} F(\gamma-\beta, \gamma-\alpha, \gamma-\alpha-\beta+1, 1-x) . \tag{175}$$

If $\gamma=\alpha+\beta$, then the two last integrals coincide and an integral of the form

$$y_5 = F(\alpha, \beta, 1, 1-x) \log(1-x) + Z(1-x) \tag{176}$$

will appear instead of the integral given by equation 175. The function $Z(1-x)$ is developable into a Taylor series in powers of $1-x$.

For equation 171 the general integral can be written in the form

$$X = MF\left(\frac{1}{12}, \frac{5}{12}, \frac{1}{2}, \xi\right) + N\xi^{1/2} F\left(\frac{7}{12}, \frac{11}{12}, \frac{3}{2}, \xi\right) , \tag{177}$$

or in the form

$$X = C_1 F\left(\frac{1}{12}, \frac{5}{12}, 1, 1-\xi\right) + C_2 \left[F\left(\frac{1}{12}, \frac{5}{12}, 1, 1-\xi\right)\log(1-\xi) + Z(1-\xi)\right] , \tag{178}$$

where $\xi=1$ is a critical logarithmic point.

However, for $\xi=1$ we have, from equation 170, $J=\infty$ and, consequently, $\Delta=0$. Thus, the case $\xi=1$ corresponds to the case of a double root of the equation $4x^3-g_2x-g_3=0$; and these roots become

$$e_1 = +\sqrt{\frac{g_2}{3}} \tag{179}$$

$$e_2 = e_3 = -\frac{1}{2}\sqrt{\frac{g_2}{3}} . \tag{180}$$

Substituting these roots into the formula for the real semi-period

$$\omega = \frac{1}{\sqrt{e_1 - e_3}} \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - k^2 \sin^2 \phi}} , \qquad k^2 = \frac{e_2 - e_3}{e_1 - e_3} ,$$

we obtain

$$\omega = \frac{\pi}{\sqrt[4]{12 g_2}}$$

or

$$X = \omega \sqrt[4]{g_2} = \frac{\pi}{\sqrt[4]{12}} \quad \text{for } \xi = 1 . \tag{181}$$

Thus the value $\xi=1$ cannot be a logarithmic singularity of X and, consequently, the value of C_2 in equation 178 must be zero. The integral of equation 171, which is of interest to us can be written either in the form given by equation 177 or in the form

$$X = C_1 F\left(\frac{1}{12}, \frac{5}{12}, 1, 1-\xi\right) . \tag{182}$$

Putting $\xi=1$ and taking equation 181 into account, we obtain

$$\omega \sqrt[4]{g_2} = \frac{\pi}{\sqrt[4]{12}} F\left(\frac{1}{12}, \frac{5}{12}, 1, 1-\xi\right) . \tag{183}$$

A linear relation must exist between the two expressions for X (as given by equation 177 and 183), which will help us to determine the values of the constants M and N. We can use the following relation between three solutions of the hypergeometric equation:

$$F(\alpha, \beta, \alpha+\beta-\gamma+1, 1-x) = \frac{\Gamma(\alpha+\beta-\gamma+1)\,\Gamma(1-\gamma)}{\Gamma(\alpha-\gamma+1)\,\Gamma(\beta-\gamma+1)} F(\alpha, \beta, \gamma, x)$$

$$+ \frac{\Gamma(\alpha+\beta+1-\gamma)\,\Gamma(\gamma-1)}{\Gamma(\alpha)\,\Gamma(\beta)} x^{1-\gamma} F(\alpha-\gamma+1, \beta-\gamma+1, 2-\gamma, x) , \tag{184}$$

which is the form given by Erdély et al.[9] By inserting into

$$\Gamma(z)\,\Gamma(1-z) = \frac{\pi}{\sin \pi z}$$

first, $z=\gamma$ and then $z=\gamma-1$, we obtain

$$\Gamma(\gamma-1) = -\frac{\Gamma(\gamma)\,\Gamma(1-\gamma)}{\Gamma(2-\gamma)} .$$

Now the relation of equation 184 can be written:

$$F(\alpha,\ \beta,\ \alpha+\beta-\gamma+1,\ 1-x) = \frac{\Gamma(\alpha+\beta-\gamma+1)\,\Gamma(1-\gamma)}{\Gamma(\alpha-\gamma+1)\,\Gamma(\beta-\gamma+1)} F(\alpha,\ \beta,\ \gamma,\ x)$$

$$- \frac{\Gamma(\alpha+\beta+1-\gamma)\,\Gamma(\gamma)\,\Gamma(1-\gamma)}{\Gamma(2-\gamma)\,\Gamma(\alpha)\,\Gamma(\beta)}\, x^{1-\gamma}\, F(\alpha-\gamma+1,\ \beta-\gamma+1,\ 2-\gamma,\ x) . \quad (185)$$

Setting $\alpha=\frac{1}{12}$, $\beta=\frac{5}{12}$, $\gamma=\frac{1}{2}$, $x=\xi$ in equation 185, we obtain

$$F\left(\tfrac{1}{12},\ \tfrac{5}{12},\ 1,\ 1-\xi\right) = \frac{\Gamma(1)\,\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{7}{12}\right)\Gamma\left(\frac{11}{12}\right)}\ F\left(\tfrac{1}{12},\ \tfrac{5}{12},\ \tfrac{1}{2},\ \xi\right)$$

$$- \sqrt{\xi}\,\frac{\Gamma^2\left(\frac{1}{2}\right)\,\Gamma(1)}{\Gamma\left(\frac{3}{2}\right)\,\Gamma\left(\frac{1}{12}\right)\,\Gamma\left(\frac{5}{12}\right)}\ F\left(\tfrac{7}{12},\ \tfrac{11}{12},\ \tfrac{3}{2},\ \xi\right) . \quad (186)$$

Putting first $n=3$, $z=-\frac{1}{12}$, and then $n=3$, $z=-\frac{1}{4}$ into the Gaussian formula

$$\frac{(2\pi)^{(n-1)/2}}{\sqrt{n}} = \frac{n^{nz}\,\Gamma(z+1)\,\Gamma\left(z+1-\frac{1}{n}\right)\cdot\cdot\cdot\,\Gamma\left(z+1-\frac{n-1}{n}\right)}{\Gamma(nz+1)} ,$$

we obtain

$$\Gamma\left(\tfrac{11}{12}\right)\,\Gamma\left(\tfrac{7}{12}\right) = \frac{2\pi}{\sqrt[4]{3}}\,\frac{\Gamma\left(\frac{3}{4}\right)}{\Gamma\left(\frac{1}{4}\right)},$$

$$\Gamma\left(\tfrac{5}{12}\right)\,\Gamma\left(\tfrac{1}{12}\right) = 2\pi\,\sqrt[4]{3}\,\frac{\Gamma\left(\frac{1}{4}\right)}{\Gamma\left(\frac{3}{4}\right)} .$$

We also have

$$\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}$$

and

$$\Gamma\left(\tfrac{3}{2}\right) = \tfrac{1}{2}\sqrt{\pi} .$$

Substituting these values of gamma-function combinations into equation 186, we obtain

$$F\left(\tfrac{1}{12},\ \tfrac{5}{12},\ 1,\ 1-\xi\right) = \frac{\sqrt[4]{3}}{2\sqrt{\pi}}\,\frac{\Gamma\left(\frac{1}{4}\right)}{\Gamma\left(\frac{3}{4}\right)}\,F\left(\tfrac{1}{12},\ \tfrac{5}{12},\ \tfrac{1}{2},\ \xi\right) - \frac{1}{\sqrt{\pi}\,\sqrt[4]{3}}\,\frac{\Gamma\left(\frac{3}{4}\right)}{\Gamma\left(\frac{1}{4}\right)}\,F\left(\tfrac{7}{12},\ \tfrac{11}{12},\ \tfrac{3}{2},\ \xi\right)\sqrt{\xi} . \quad (187)$$

Multiplying both sides of the last equation by $\pi/(12)^{1/4}$ and taking equation 183 into account we have

$$X = \frac{\sqrt{\pi}}{2\sqrt{2}} \frac{\Gamma\left(\frac{1}{4}\right)}{\Gamma\left(\frac{3}{4}\right)} F\left(\frac{1}{12}, \frac{5}{12}, \frac{1}{2}, \xi\right) - \frac{\sqrt{\pi}}{\sqrt{6}} \frac{\Gamma\left(\frac{3}{4}\right)}{\Gamma\left(\frac{1}{4}\right)} \sqrt{\xi}\, F\left(\frac{7}{12}, \frac{11}{12}, \frac{3}{2}, \xi\right) . \tag{188}$$

In order to simplify the last equation we shall make use of the B-function:

$$B(u, v) = 2 \int_0^1 x^{2u-1} \left(1 - x^2\right)^{v-1} dx , \tag{189}$$

and

$$B(u, v) = \frac{\Gamma(u)\, \Gamma(v)}{\Gamma(u+v)} . \tag{190}$$

Setting, in the last equations, first

$$u = \frac{1}{4}, \quad v = \frac{1}{2}$$

and then

$$u = \frac{3}{4}, \quad v = \frac{1}{2},$$

we have

$$B\left(\frac{1}{4}, \frac{1}{2}\right) = 2 \int_0^1 x^{-1/2} \left(1 - x^2\right)^{-1/2} dx = \frac{\Gamma\left(\frac{1}{4}\right) \Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{3}{4}\right)}$$

$$B\left(\frac{3}{4}, \frac{1}{2}\right) = 2 \int_0^1 x^{1/2} \left(1 - x^2\right)^{-1/2} dx = \frac{\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{3}{4}\right)}{\Gamma\left(\frac{5}{4}\right)} .$$

Taking into account

$$\frac{1}{4} \Gamma\left(\frac{1}{4}\right) = \Gamma\left(\frac{5}{4}\right)$$

and

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

and introducing a new variable $y = x^{1/2}$ instead of x, we have

$$\frac{\sqrt{\pi}}{2} \frac{\Gamma\left(\frac{1}{4}\right)}{\Gamma\left(\frac{3}{4}\right)} = 2 \int_0^1 \frac{dx}{\sqrt{1 - x^4}} ,$$

$$\sqrt{\pi}\, \frac{\Gamma\left(\frac{3}{4}\right)}{\Gamma\left(\frac{1}{4}\right)} = \int_0^1 \frac{x^2\, dx}{\sqrt{1 - x^4}} .$$

And equation 188 takes the form

$$X\sqrt{2} = 2A_1 F\left(\frac{1}{12}, \frac{5}{12}, \frac{1}{2}, \xi\right) - B_1 \sqrt{\frac{\xi}{3}} F\left(\frac{7}{12}, \frac{11}{12}, \frac{3}{2}, \xi\right), \tag{191}$$

where

$$A_1 = \int_0^1 \frac{dx}{\sqrt{1-x^4}} = 1.311028777146 \cdots,$$

$$B_1 = \int_0^1 \frac{x^2\,dx}{\sqrt{1-x^4}} = 0.599070117367 \cdots.$$

Halphen used $\Psi(\xi) = X\sqrt{2}$ instead of X. From equation 191 we have

$$\Psi(\xi) = 2A_1 F\left(\frac{1}{12}, \frac{5}{12}, \frac{1}{2}, \xi\right) - B_1 \sqrt{\frac{\xi}{3}} F\left(\frac{7}{12}, \frac{11}{12}, \frac{3}{2}, \xi\right), \tag{192}$$

and from equation 183

$$\Psi(\xi) = C_1 F\left(\frac{1}{12}, \frac{5}{12}, 1, 1-\xi\right), \tag{193}$$

where $C_1 = \pi(3)^{1/4}$. Goriachev used both forms of $\Psi(\xi)$.

Equation 192 is convenient if $\xi \leq \frac{1}{2}$ and equation 193 is preferable if $\xi \geq \frac{1}{2}$. However, remembering that a hypergeometric series converges rather slowly, we shall find it more convenient to transform equation 193 to a fast convergent form and to use the transformed series throughout the interval $0 \leq \xi \leq 1$.

By applying the Goursat transformation[6]

$$F\left(\alpha, \beta, \alpha+\beta+\frac{1}{2}, z\right) = \left(\frac{1+\sqrt{1-z}}{2}\right)^{-2\alpha} F\left(2\alpha, \alpha-\beta+\frac{1}{2}, \alpha+\beta+\frac{1}{2}, -\frac{1-\sqrt{1-z}}{1+\sqrt{1-z}}\right)$$

to equation 193, we obtain

$$\Psi(\xi) = \frac{\pi}{4\sqrt{3}} \sqrt[6]{\frac{2}{1+\sqrt{\xi}}}\, F\left(\frac{1}{6}, \frac{1}{6}, 1, -\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}\right). \tag{194}$$

The hypergeometric series which appears in the last equation is an alternating series and its convergence can be sped up considerably by applying the Euler summability process.

The general formula of the Euler summability process as applied to the series $a_0+a_1+a_2+a_3+\cdots$ can be written in the form[10]:

$$\sum_{k=0}^{\infty} a_k = \sum_{k=0}^{N} a_k + \lim_{m\to\infty}\left[\sum_{j=1}^{m} a_{N+j} \sum_{p=j}^{m} \frac{1}{2^{p+1}} \binom{p}{j}\right]. \tag{195}$$

For the hypergeometric series

$$F(\alpha, \beta, \gamma, x) = \sum_{k=0}^{\infty} \frac{(\alpha, k)(\beta, k)}{(1, k)(\gamma, k)} x^k, \qquad (m, k) = m(m+1)\cdots(m+k-1),$$

the formula (equation 195) takes the form:

$$F(\alpha, \beta, \gamma, x) = \sum_{k=0}^{N} \frac{(\alpha, k)(\beta, k)}{(1, k)(\gamma, k)} x^k + \lim_{m\to\infty}\left[\sum_{j=1}^{m} \frac{(\alpha, N+j)(\beta, N+j)}{(1, N+j)(\gamma, N+j)} x^{N+j} \sum_{p=j}^{m} \frac{1}{2^{p+1}} \binom{p}{j}\right].$$

Here, we have $\alpha=\frac{1}{6}$, $\beta=\frac{1}{6}$ and $\gamma=1$. A high degree of approximation is already obtained by setting $N=3$ and $m=19$. The final result is:

$$\omega \sqrt[4]{4g_2} = \Psi(\xi) = \left(\frac{2}{1+\sqrt{\xi}}\right)^{1/6} \times$$

$$\begin{aligned} \times \big(& + 2.3870942 \\ & - 0.0663082\, w \\ & + 0.0225632\, w^2 \\ & - 0.0117691\, w^3 \\ & + 0.0073743\, w^4 \\ & - 0.0051060\, w^5 \\ & + 0.0037250\, w^6 \\ & - 0.0027325\, w^7 \\ & + 0.0019070\, w^8 \\ & - 0.0011936\, w^9 \\ & + 0.0006337\, w^{10} \\ & - 0.0002710\, w^{11} \\ & + 0.0000884\, w^{12} \\ & - 0.0000205\, w^{13} \\ & + 0.0000030\, w^{14} \\ & - 0.0000002\, w^{15} \big), \end{aligned} \quad (196)$$

where

$$w = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}},$$

and the convergence is fast. The program written by A. J. Smith, Jr., of the Theoretical Division, Goddard Space Flight Center makes use of equations 196 and 207.*

* This program will be discussed in the following part 2 of this treatise, by A. J. Smith, Jr., NASA TR R-194.

A comparison of Goriachev's computations for the minor planet (1) Ceres (done on the basis of equations 192 and 193) with the results obtained by Smith using equation 196 shows a complete agreement between the two methods of computation. The coefficient B (equation 89)

$$B = \frac{2\omega}{\pi g_2},$$

which appears in the expression for the dyadic Φ (equation 113) and in the expressions for the disturbing force F_0 (equations 114b and 134), can now be put into the final form. We have, taking equation 196 into account,

$$B = \frac{\sqrt{2}}{\pi g_2 \sqrt[4]{g_2}} \Psi(\xi). \quad (197)$$

We have still to form a convenient expression for the coefficient A (equation 88)

$$A = \frac{16}{\pi\Delta}\left(\frac{3}{2}\frac{g_3}{g_2}\omega - \eta\right).$$

We have

$$\xi = \frac{27 g_3^2}{g_2^3}, \quad (198)$$

$$D\xi = D\frac{27 g_3^2}{g_2^3} = D\left(1-\frac{1}{J}\right) = \frac{1}{J^2} DJ; \quad (199)$$

and by considering equation 153 we deduce

$$D\xi = 4\sqrt{3}\,\Delta^{1/6} J^{-4/3} (J-1)^{1/2} \quad (200)$$

from the last equation. Eliminating J and $J-1$ in favor of g_2 and g_3 by means of $J=g_2^3/\Delta$ and $J-1=27g_3^2/\Delta$, we deduce from equation 200 that

$$D\xi = \frac{36 g_3}{g_2^4}\Delta. \quad (201)$$

We also have

$$D\left(g_2^{1/4}\right) = \frac{3 g_3}{g_2 \sqrt[4]{g_2}} \quad (202)$$

and $D\omega = -2\eta$ (equation 150a).

Let us now apply the operator D to $\omega(4g_2)^{1/4}=\Psi(\xi)$. Taking equations 201, 202, and 150a into account, we deduce that

$$-2\eta \sqrt[4]{4g_2} + \omega \frac{3g_3}{g_2} \sqrt[4]{4g_2} = \frac{36 g_3}{g_2^4} \cdot \Delta \cdot \Psi'(\xi)$$

or

$$\frac{16}{\Delta}\left(\frac{3}{2}\frac{g_3}{g_2}\omega - \eta\right) = \frac{144\sqrt{2}}{\sqrt[4]{g_2}} \frac{1}{g_2^4} g_3 \Psi'(\xi) ,$$

and thus

$$A = \frac{144 g_3 \sqrt{2}}{\pi g_2^4 \sqrt[4]{g_2}} \Psi'(\xi) . \tag{203}$$

Now by applying the formula

$$\frac{d}{dx} F(\alpha, \beta, \gamma, x) = \frac{\alpha\beta}{\gamma} F(\alpha+1, \beta+1, \gamma+1, x)$$

to $\Psi(\xi)$ as given by equation 193 we obtain a value of $\Psi'(\xi)$ which may be substituted into equation 203 to yield

$$A = -\frac{5g_3\sqrt{2}}{g_2^4 \sqrt[4]{3g_2}} F\left(\frac{13}{12}, \frac{17}{12}, 2, 1-\xi\right) . \tag{204}$$

Taking the equation

$$\xi = \frac{J-1}{J} = \frac{27 g_3^2}{g_2^3}$$

into account, we deduce from equation 204, by eliminating g_3 in favor of ξ and g_2, that

$$A = -\frac{\sqrt[4]{12}}{9} \cdot \frac{\sqrt[4]{g_2}}{g_2^3} \sqrt{\xi} F\left(\frac{13}{12}, \frac{17}{12}, 2, 1-\xi\right) . \tag{205}$$

By applying the Goursat transformation

$$F\left(\alpha, \beta, \alpha+\beta-\frac{1}{2}, z\right)$$

$$= (1-z)^{-1/2} \left(\frac{1+\sqrt{1-z}}{2}\right)^{1-2\alpha} F\left(2\alpha-1, \alpha-\beta+\frac{1}{2}, \alpha+\beta-\frac{1}{2}, -\frac{1-\sqrt{1-z}}{1+\sqrt{1-z}}\right)$$

to the hypergeometric series in the right side of equation 205, we obtain, finally,

$$A = -\frac{5}{\sqrt[4]{3}}\left(\frac{2}{1+\sqrt{\xi}}\right)^{7/6} F\left(\frac{1}{6}, \frac{7}{6}, 2, -w\right) \frac{\sqrt{6}}{9} \frac{\sqrt[4]{g_2}}{g_2^3} . \tag{206}$$

The Euler summability process can be applied to the hypergeometric series of the last equation. We have

$$-\frac{5}{\sqrt[4]{3}}F\left(\frac{1}{6},\frac{7}{6},2,-w\right) = -3.7991784 + 0.3693646\,w - 0.1556119\,w^2 + 0.0889726\,w^3 - 0.0586828\,w^4 + 0.0419870\,w^5 - 0.0313364\,w^6 + 0.0233758\,w^7 - 0.0165247\,w^8 + 0.0104483\,w^9 - 0.0055933\,w^{10} + 0.0024083\,w^{11} - 0.0007898\,w^{12} + 0.0001837\,w^{13} - 0.0000268\,w^{14} + 0.0000018\,w^{15} \tag{207}$$

where, as before, $w=(1-\xi^{1/2})/(1+\xi^{1/2})$. Equation 207 was used instead of Goriachev's tables in the actual computations.

EQUATIONS FOR SECULAR VARIATIONS OF ELEMENTS

Let $fm'S$, $fm'T$, $fm'Z$ be the radial, the tangential, and the normal components respectively of the disturbing force **F**. We have[11] for the variation of elliptic elements a, e, $\pi=\omega+\Omega$, $L=g+\pi$, i, and Ω

$$\frac{da}{dt} = \frac{2m'na^2}{M+m}\left(S\frac{ae}{\sqrt{1-e^2}}\sin v + T\frac{a^2}{r}\sqrt{1-e^2}\right);$$

$$\frac{de}{dt} = \frac{m'}{M+m}\cdot\frac{na(1-e^2)}{e}\left(S\frac{ae}{\sqrt{1-e^2}}\sin v + T\frac{a^2}{r}\sqrt{1-e^2}\right) - \frac{m'}{M+m}\frac{na\sqrt{1-e^2}}{e}Tr;$$

$$\frac{d\pi}{dt} = \frac{m'}{M+m}\cdot\frac{na\sqrt{1-e^2}}{e}\left[-S a\cos v + Ta\left(1+\frac{1}{1-e^2}\cdot\frac{r}{a}\right)\sin v\right] + 2\sin^2\frac{i}{2}\frac{d\Omega}{dt};$$

$$\frac{dL}{dt} = -\frac{m'}{M+m}2nar S + \left(1-\sqrt{1-e^2}\right)\frac{d\pi}{dt} + 2\sqrt{1-e^2}\sin^2\frac{i}{2}\frac{d\Omega}{dt};$$

$$\frac{di}{dt} = \frac{m'}{M+m}\cdot\frac{na}{\sqrt{1-e^2}}Zr\cos(v+\omega);$$

$$\sin i\frac{d\Omega}{dt} = \frac{m'}{M+m}\cdot\frac{na}{\sqrt{1-e^2}}Zr\sin(v+\omega).$$

Let S_0, T_0, Z_0 be the values of S, T, Z averaged over the orbit of the disturbing body:

$$S_0 = \frac{1}{2\pi}\int_0^{2\pi} S\,dg' ,$$

$$T_0 = \frac{1}{2\pi}\int_0^{2\pi} T\,dg' ,$$

$$Z_0 = \frac{1}{2\pi}\int_0^{2\pi} Z\,dg' .$$

Averaging the equations for the variation of elements with respect to g and g' and taking the equations

$$r\cos v = a\cos E - ae ,$$

$$r\sin v = a\sqrt{1-e^2}\sin E ,$$

$$r = a - ae\cos E ,$$

$$E - e\sin E = g ,$$

$$dg = \frac{r}{a}\,dE$$

into account, we deduce the following equations for secular variations of elliptic elements:

$$\frac{da}{dt} = \frac{2m'na^3}{M+m}\cdot\frac{1}{2\pi}\int_0^{2\pi}\left(e S_0\sin E + T_0\sqrt{1-e^2}\right)dE ;$$

$$\frac{de}{dt} = \frac{m'na^2\sqrt{1-e^2}}{M+m}\cdot\frac{1}{2\pi}\int_0^{2\pi}\left[S_0\sqrt{1-e^2}\sin E + T_0\left(-\frac{3}{2}e + 2\cos E - \frac{1}{2}e\cos 2E\right)\right]dE ;$$

$$\frac{d\pi}{dt} = \frac{m'na^2\sqrt{1-e^2}}{(M+m)e}\cdot\frac{1}{2\pi}\int_0^{2\pi}\left[-S_0(\cos E - e) + T_0\left(1 + \frac{1}{1-e^2}\cdot\frac{r}{a}\right)\sqrt{1-e^2}\sin E\right]dE + 2\sin^2\frac{i}{2}\frac{d\Omega}{dt} ;$$

$$\frac{dL}{dt} = -\frac{2m'na^2}{M+m}\cdot\frac{1}{2\pi}\int_0^{2\pi} S_0\frac{r^2}{a^2}\,dE + \frac{e}{1+\sqrt{1-e^2}}\left(e\frac{d\pi}{dt}\right) + 2\sqrt{1-e^2}\sin^2\frac{i}{2}\frac{d\Omega}{dt} ;$$

$$\frac{di}{dt} = \frac{m'na^2}{(M+m)\sqrt{1-e^2}}\cdot\frac{1}{2\pi}\int_0^{2\pi} Z_0\frac{r}{a}\left[(\cos E - e)\cos\omega - \sqrt{1-e^2}\sin E\sin\omega\right]dE ;$$

$$\sin i\frac{d\Omega}{dt} = \frac{m'na^2}{(M+m)\sqrt{1-e^2}}\cdot\frac{1}{2\pi}\int_0^{2\pi} Z_0\frac{r}{a}\left[(\cos E - e)\sin\omega + \sqrt{1-e^2}\sin E\cos\omega\right]dE .$$

The values of S_0, T_0, Z_0 are computed analytically by using the formulas developed in the previous two sections. The integrals with respect to E are computed numerically by giving E a set of particular values conveniently distributed over the orbit of the disturbed body.

The secular variation of da/dt is zero in this theory and in the process of computation the smallness of da/dt will determine the range of validity of the theory and, at the same time, serve as a check of the accuracy of the computation.

CONCLUSION

The collection of formulas given in appendix A was programed for the actual computation of long range effects in the motion of artificial satellites and minor planets using step by step integration. Halphen's method can be especially useful when near-resonance conditions arise. In this case difficulties are caused by the presence of a small divisor of the form $i\dot{\omega}+i'\dot{\omega}'+j\dot{\Omega}+j'\dot{\Omega}'$ if the problem is treated analytically.

In the case of an artificial satellite Halphen's method might give information on the long range effects and the stability of orbit over the interval of approximately 15–20 years. In the case of minor planets, it can supply information about the long range ("secular") effects in the elements over intervals of hundreds of thousands of years; the integration step can be taken to be 100–500 years. We assume that no sharp commensurability between mean motions of the disturbed and disturbing bodies exists. The secular variations of the elements of the disturbing bodies are also taken into consideration.

ACKNOWLEDGMENT

The author would like to take the opportunity to express his gratitude to Mr. Arthur J. Smith, Jr., who programed the theory and without whose generous assistance the numerical part of the work could never have been completed.

REFERENCES

1. Brown, E. W., and Shook, C. A., "Planetary Theory," Cambridge (England): The University Press, 1933, p. 159.
2. Gauss, K. F., "Determinatio attractionis quam in punctum quodvis positionis datae exerceret planeta si eius massa per totam orbitam ratione temporis quo singulae partes describuntur uniformiter esset dispertita," (1818) in: His *Werke*, Göttingen: Königliche Gesellschaft der Wissenschaften, 1866, vol. 3, pp. 331–335.
3. Halphen, G. H., "Traité des Fonctions Elliptiques et de Leurs Applications," vol. 2, Paris: Gauthier–Villars et Fils, 1888.
4. Musen, P., "On the Long-Period Lunisolar Effect in the Motion of the Artificial Satellite," *J. Geophys. Res.* **66**(6): 1659–1665, June 1961.
5. Goriachev, N. N., "Halphen's Method for the Computation of Secular Perturbations," University of Tomsk, 1937 (in Russian).
6. Goursat, É., "Sur l'Équation Différentielle Linéaire qui Admet pour Intégrale la Série Hypergéométrique," *Annales Scientifiques de l'École Normale Supérieure, Supplément* Ser. 2, **10:** 3–142, 1881.
7. Gibbs, J. W., "Vector Analysis," New Haven: Yale University Press, 1901.
8. Hancock, H., "Lectures on the Theory of Elliptic Functions: Analysis," New York: Dover, 1958.
9. Bateman Manuscript Project, California Institute of Technology, "Higher Transcendental Functions," vol. 1, New York: McGraw-Hill, 1953.
10. Knopp, K., "Infinite Sequences and Series," New York: Dover, 1956, p. 144.
11. Brown, E. W., "An Introductory Treatise on the Lunar Theory," Cambridge (England): The University Press, 1896, p. 63.

Appendix A

COLLECTION OF FORMULAS

I. Elements of the disturbed planet:

$$g_0,\ a,\ n,\ i,\ \omega,\ \Omega,\ e,\ b.$$

Elements of the disturbing planet:

$$g_0',\ a',\ n',\ i',\ \omega'\ \ \Omega',\ e',\ b',\ .$$

II.

$$\begin{bmatrix} P_x & Q_x & R_x \\ P_y & Q_y & R_y \\ P_z & Q_z & R_z \end{bmatrix} = \begin{bmatrix} +\cos\Omega & -\sin\Omega & 0 \\ +\sin\Omega & +\cos\Omega & 0 \\ 0 & 0 & +1 \end{bmatrix} \cdot \begin{bmatrix} +1 & 0 & 0 \\ 0 & +\cos i & -\sin i \\ 0 & +\sin i & +\cos i \end{bmatrix} \cdot \begin{bmatrix} +\cos\omega & -\sin\omega & 0 \\ +\sin\omega & +\cos\omega & 0 \\ 0 & 0 & +1 \end{bmatrix}$$

$$\begin{bmatrix} P_x' & Q_x' & R_x' \\ P_y' & Q_y' & R_y' \\ P_z' & Q_z' & R_z' \end{bmatrix} = \begin{bmatrix} +\cos\Omega' & -\sin\Omega' & 0 \\ +\sin\Omega' & +\cos\Omega' & 0 \\ 0 & 0 & +1 \end{bmatrix} \cdot \begin{bmatrix} +1 & 0 & 0 \\ 0 & +\cos i' & -\sin i' \\ 0 & +\sin i' & +\cos i' \end{bmatrix} \cdot \begin{bmatrix} +\cos\omega' & -\sin\omega' & 0 \\ +\sin\omega' & +\cos\omega' & 0 \\ 0 & 0 & +1 \end{bmatrix} .$$

III. For example, for $E=0°, 10°, 20°, \cdots, 350°$:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} P_x' & P_y' & P_z' \\ Q_x' & Q_y' & Q_z' \\ R_x' & R_y' & R_z' \end{bmatrix} \cdot \begin{bmatrix} P_x & Q_x & R_x \\ P_y & Q_y & R_y \\ P_z & Q_z & R_z \end{bmatrix} \cdot \begin{bmatrix} a(\cos E - e) \\ a\sqrt{1-e^2}\,\sin E \\ 0 \end{bmatrix};$$

$$\alpha = x + e'a', \quad \beta = y, \quad \gamma = z.$$

IV.

$$k_1 = \alpha^2 + \beta^2 + \gamma^2 - (a'^2 + b'^2)$$

$$k_2 = a'^2 b'^2 - b'^2 \alpha^2 - a'^2 \beta^2 - (a'^2 + b'^2)\gamma^2$$

$$k_3 = a'^2 b'^2 \gamma^2$$

$$g_2 = \frac{4}{3}(k_1^2 - 3k_2)$$

$$g_3 = \frac{4}{27}(2k_1^3 - 9k_1 k_2 + 27k_3)$$

$$\xi = \frac{27 g_3^2}{g_2^3}$$

$$k_4 = 9k_3 - k_1 k_2 ,$$

$$k_5 = k_1 (k_1 k_2 - 3k_3) - 2k_2^2 .$$

V.

$$A = \frac{\sqrt{6}\sqrt[4]{g_2}}{9 g_2^3} \cdot \frac{144}{\pi}\sqrt{\xi}\,\Psi'(\xi)$$

$$B = \frac{\sqrt{2}}{\pi g_2 \sqrt[4]{g_2}}\,\Psi(\xi)$$

$$w = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}$$

$$\Psi(\xi) = \left(\frac{2}{1+\sqrt{\xi}}\right)^{1/6} \times$$
$$\times (+ 2.3870942$$
$$- 0.0663082\ w$$
$$+ 0.0225632\ w^2$$
$$- 0.0117691\ w^3$$
$$+ 0.0073743\ w^4$$
$$- 0.0051060\ w^5$$
$$+ 0.0037250\ w^6$$
$$- 0.0027325\ w^7$$
$$+ 0.0019070\ w^8$$
$$- 0.0011936\ w^9$$
$$+ 0.0006337\ w^{10}$$
$$- 0.0002710\ w^{11}$$
$$+ 0.0000884\ w^{12}$$
$$- 0.0000205\ w^{13}$$
$$+ 0.0000030\ w^{14}$$
$$- 0.0000002\ w^{15})$$

$$\frac{144}{\pi}\sqrt{\xi}\,\Psi'(\xi) = \left(\frac{2}{1+\sqrt{\xi}}\right)^{7/6} \times$$
$$\times (- 3.7991784$$
$$+ 0.3693646\ w$$
$$- 0.1556119\ w^2$$
$$+ 0.0889726\ w^3$$
$$- 0.0586828\ w^4$$
$$+ 0.0419870\ w^5$$
$$- 0.0313364\ w^6$$
$$+ 0.0233758\ w^7$$
$$- 0.0165247\ w^8$$
$$+ 0.0104483\ w^9$$
$$- 0.0055933\ w^{10}$$
$$+ 0.0024083\ w^{11}$$
$$- 0.0007898\ w^{12}$$
$$+ 0.0001837\ w^{13}$$
$$- 0.0000268\ w^{14}$$
$$+ 0.0000018\ w^{15})$$

VI.

$$a_{11} = k_4 \left(\alpha^2 - a'^2\right) + k_5 + \frac{3}{2} \frac{g_2 k_3}{a'^2} ,$$

$$a_{22} = k_4 \left(\beta^2 - b'^2\right) + k_5 + \frac{3}{2} \frac{g_2 k_3}{b'^2} ,$$

$$a_{33} = k_4 \gamma^2 + k_5 + \frac{3}{2} g_2 \left(\alpha^2 b'^2 + \beta^2 a'^2 - a'^2 b'^2\right) ,$$

$$a_{12} = a_{21} = k_4 \alpha\beta ,$$

$$a_{23} = a_{32} = k_4 \beta\gamma - \frac{3}{2} g_2 a'^2 \beta\gamma ,$$

$$a_{31} = a_{13} = k_4 \gamma\alpha - \frac{3}{2} g_2 b'^2 \gamma\alpha ,$$

$$a'_{11} = \alpha^2 - a'^2 - \frac{1}{3} k_1 ,$$

$$a'_{22} = \beta^2 - b'^2 - \frac{1}{3} k_1 ,$$

$$a'_{33} = \gamma^2 - \frac{1}{3} k_1 ,$$

$$a'_{12} = a'_{21} = \alpha\beta ,$$

$$a'_{23} = a'_{32} = \beta\gamma ,$$

$$a'_{31} = a'_{13} = \gamma\alpha ,$$

$$A_{ij} = a_{ij} A + a'_{ij} B \qquad (i, j = 1, 2, 3)$$

$$A_{ij} = A_{ji} .$$

VII.

$$\begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} = -2 \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} .$$

VIII.

$$r = a(1 - e \cos E)$$

$$\alpha_1 = \frac{x}{r},$$

$$\beta_1 = \frac{y}{r},$$

$$\gamma_1 = \frac{z}{r},$$

$$\begin{bmatrix} \alpha_3 \\ \beta_3 \\ \gamma_3 \end{bmatrix} = \begin{bmatrix} P_x' & P_y' & P_z' \\ Q_x' & Q_y' & Q_z' \\ R_x' & R_y' & R_z' \end{bmatrix} \cdot \begin{bmatrix} R_x \\ R_y \\ R_z \end{bmatrix}$$

$$\alpha_2 = \gamma_1 \beta_3 - \beta_1 \gamma_3$$

$$\beta_2 = \alpha_1 \gamma_3 - \gamma_1 \alpha_3$$

$$\gamma_2 = \beta_1 \alpha_3 - \alpha_1 \beta_3 .$$

IX.

$$\begin{bmatrix} S_0 \\ T_0 \\ Z_0 \end{bmatrix} = \begin{bmatrix} \alpha_1 & \beta_1 & \gamma_1 \\ \alpha_2 & \beta_2 & \gamma_2 \\ \alpha_3 & \beta_3 & \gamma_3 \end{bmatrix} \cdot \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} .$$

X.

$$\frac{da}{dt} = \frac{2m' na^3}{M+m} \cdot \frac{1}{2\pi} \int_0^{2\pi} \left(e S_0 \sin E + T_0 \sqrt{1-e^2} \right) dE ,$$

$$\frac{de}{dt} = \frac{m' na^2 \sqrt{1-e^2}}{M+m} \cdot \frac{1}{2\pi} \int_0^{2\pi} \left[S_0 \sqrt{1-e^2} \sin E + T_0 \left(-\frac{3}{2} e + 2\cos E - \frac{1}{2} e \cos 2E \right) \right] dE ,$$

$$\frac{d\pi}{dt} = \frac{m' na^2 \sqrt{1-e^2}}{(M+m)e} \cdot \frac{1}{2\pi} \int_0^{2\pi} \left[-S_0 (\cos E - e) + T_0 \left(1 + \frac{1}{1-e^2} \cdot \frac{r}{a} \right) \sqrt{1-e^2} \sin E \right] dE + 2 \sin^2 \frac{i}{2} \frac{d\Omega}{dt} ,$$

$$\frac{dL}{dt} = -\frac{2m' na^2}{M+m} \cdot \frac{1}{2\pi} \int_0^{2\pi} S_0 \cdot \frac{r^2}{a^2} dE + \frac{e}{1+\sqrt{1-e^2}} \left(e \frac{d\pi}{dt} \right) + 2 \sqrt{1-e^2} \sin^2 \frac{i}{2} \frac{d\Omega}{dt} ,$$

$$\frac{di}{dt} = \frac{m' na^2}{(M+m)\sqrt{1-e^2}} \cdot \frac{1}{2\pi} \int_0^{2\pi} Z_0 \frac{r}{a} \left[(\cos E - e) \cos\omega - \sqrt{1-e^2} \sin E \sin\omega \right] dE ,$$

$$\sin i \frac{d\Omega}{dt} = \frac{m' na^2}{(M+m)\sqrt{1-e^2}} \cdot \frac{1}{2\pi} \int_0^{2\pi} Z_0 \frac{r}{a} \left[(\cos E - e) \sin\omega + \sqrt{1-e^2} \sin E \cos\omega \right] dE .$$

ON LONG RANGE EFFECTS IN THE MOTION OF ARTIFICIAL SATELLITES

PETER MUSEN

Goddard Space Flight Center

The long range effects as caused by the moon and the sun are of primary importance for establishing the stability of highly eccentric orbits of satellites. For the time being no complete analytical theory exists which can treat such orbits. The use of Halphen's method of treating secular planetary effects was suggested for the determination of the long range lunar effects in the motion of artificial satellites.

This choice permits the numerical integration of long range lunar effects over an interval of many years. Numerous examples are computed and plotted. The long range solar effects are treated by averaging the disturbing function over the orbit of the satellite.

The effects in the motion of a 24-hour satellite caused by the ellipticity of earth's equator are treated using a resonance theory.

This paper contains results obtained by a group working in celestial mechanics in the Theoretical Division at the Goddard Space Flight Center. Consideration is given to the long range effects in the motion of artificial satellites with large eccentricities, large inclinations and large semi-major axes. The effects of the moon and the sun and the irregularities of the earth's gravitational field are considered. In treating the problem of stability of such satellites, the long range effects are of primary importance, but no purely analytical theory considering the long range effects for such extreme elements exists. One has to resort to numerical integration in order to obtain information about the stability of the orbit over a long interval of time and about the lifetime of the satellite. The methods based on the use of an unaveraged disturbing function, like Cowell or Encke, contain both the short and the long period terms, and, in the case of artificial satellites, require that the interval of integration be much less than that of the period of the satellite. The main long-range effects in the elements are produced by the long range terms in the disturbing function and by their "cross actions". The short period terms can produce long-range effects through their mutual cross actions in higher approximations, but these are very small (Brown and Shook, 1933).

For these reasons, as well as to diminish the accumulation of round off errors, it is necessary to remove the short period terms from the disturbing function or from the components of the disturbing force from the very start.

The problem thus formulated does not differ from the problem of determining the secular effects of planets and comets by means of numerical integration using the Gaussian method (1818). With the existence of modern electronic equipment such a solution of the problem has become possible.

The use of Halphen's form (1888) of the Gaussian theory was suggested by the author as a practical method for determining the long-range lunar effects in the motion of artificial satellites (Musen, 1961). The secular changes in the moon's elements as well as the perturbations in the satellite's elements caused by the earth's oblateness were included into the programing of the method for the use of IBM 7090.

Halphen's method previously was not in use, probably, because of several numerical errors which appear in the original publication. They

were all corrected by Goriachev (1937), whose name should be associated with method as well, and in its present form should justly be called the Halphen–Goriachev method. The Halphen–Goriachev method is based on simple operations with matrices. The trigonometrical transformations which are so typical to Hill's method (1901) are absent in the Halphen–Goriachev method.

The author has suggested (1961) the use of the Goursat transformation (1881) and of the E-summability process to speed up the convergence

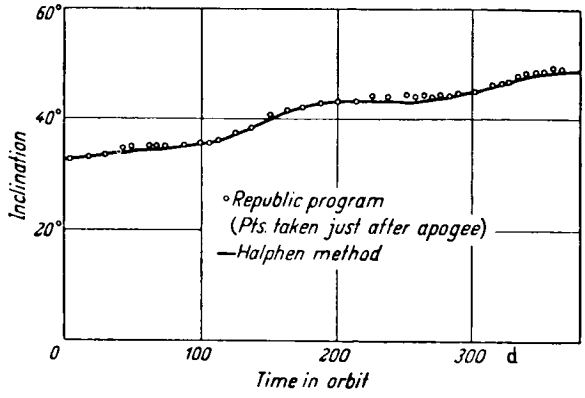


FIG. 2—IMP inclination vs. time

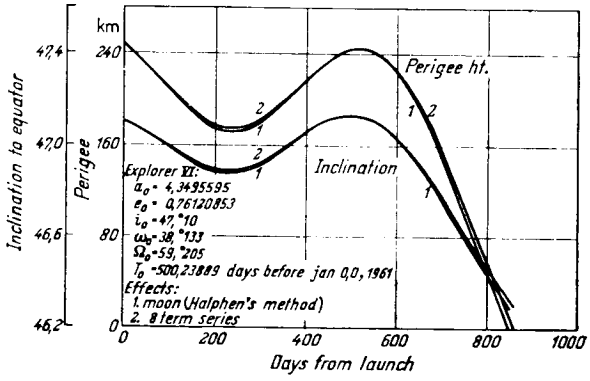


FIGURE 1.

of two hypergeometric series which appear in the Halphen–Goriachev method and to facilitate the numerical computation. All programing for the use of the electronic machine was done by A. Smith of the Theoretical Division.

A careful comparison of the Halphen–Goriachev method with some other existing methods was done before it was recommended for use on a large scale. Figure 1 shows the results of a comparison of the Halphen–Goriachev method, as applied to the determination of long-range lunar effects in the motion of Explorer VI, with the method based on the use of trigonometric series as developed by the author, Bailie, and Upton (1961). The lifetime prediction agrees with the prediction by Kozai (1959).

B. Shute of the Theoretical Division has applied the Halphen–Goriachev method extensively to the investigation of long-range stability of artificial satellites and compared the results for some satellites with numerical integration using Encke's method as programed by the Republic Aviation Corp., Farmingdale, N. Y. Figures 2 and 3 show the results obtained using both methods for the hypothetical satellite IMP ("Interplanetary Monitoring Probe"). This satellite has a semimajor axis of 22 earth radii and an eccentricity equal to 0.95 initially. The comparison clearly indicates a relative unimportance of short period terms and of their long range cross actions over an extended interval of time. One might expect that the effect of such cross action terms will not exceed approximately 1% to 2% of the perturbations obtained on the basis of the Halphen–Goriachev method, providing that no sharp commensurability of mean motions does exist.

Figures 4 and 5 show the results, also obtained by B. Shute, for IMP over an interval of 12 years. They show that for satellites of IMP type one might expect large variations of the inclination

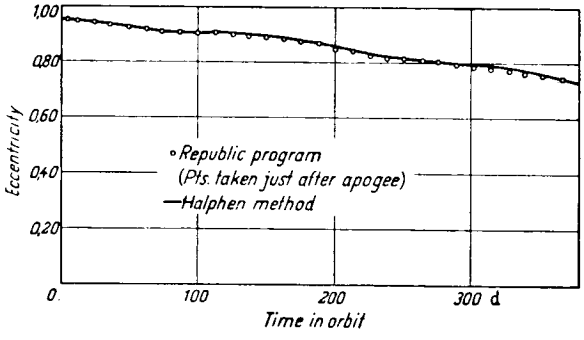


FIGURE 3.—IMP eccentricity vs. time.

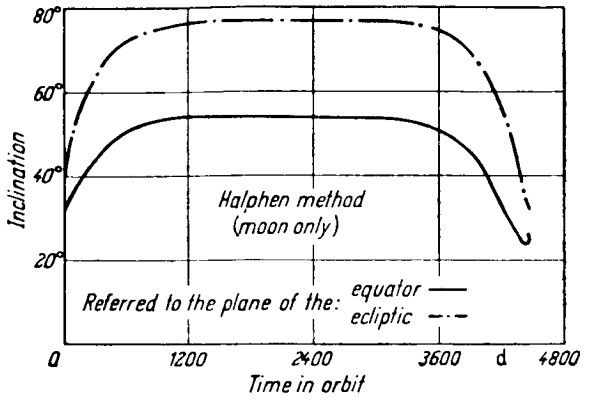


FIGURE 4.—IMP inclination vs. time, 10-year study.

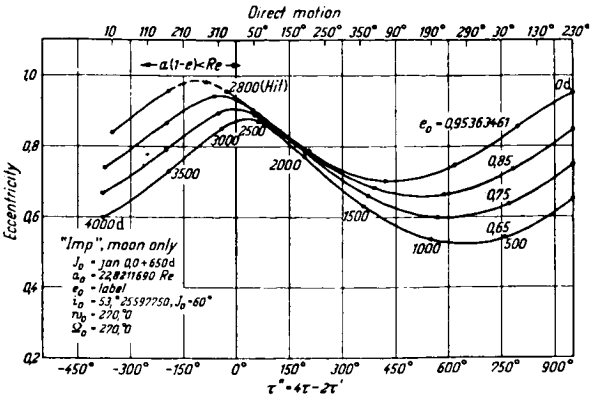


FIGURE 6.

and the eccentricity under the influence of the moon. The lifetime of IMP will be approximately 8 years for these initial conditions. Such large "secular" changes in the elements might also take place in the core of minor planets and comets, but we never will be able to observe the accumulation of such effects directly. Similar events take place in an earth-moon-satellite system in a much shorter time. In that fact lies the importance of artificial satellites for controlled experiments in the domain of "secular" perturbations and for sharpening the mathematical tools of celestial mechanics.

The Halphen–Goriachev method might be helpful in the problem of the determination of characteristic points in the phase space of orbital elements which are associated with the problem of stability.

Let τ be the argument of the perigee and $-\tau'$ be the longitude of the ascending node with respect to the moon's orbit. Figure 6 is a plot of the eccentricity versus a critical argument $4\tau-2\tau'$ for some particular values of the initial elements. The existence of a node is evident. Figures 7 and 8 give example of a focus for a direct and for a retrograde motion of an artificial satellite. The case of the retrograde motion shows more stability than the case of the direct motion. This statement can be easily proved analytically for small eccentricities and inclinations (Hirayama, 1927). In a more general case the answer can be obtained by systematic investigations based on the Halphen–Goriachev method.

Finally, the same method can be used to deduce

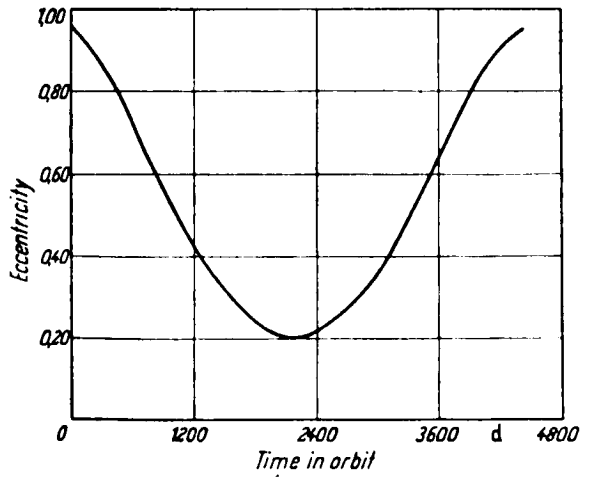


FIGURE 5.—IMP eccentricity vs. time, 10-year study. Halphen method (*Moon only*).

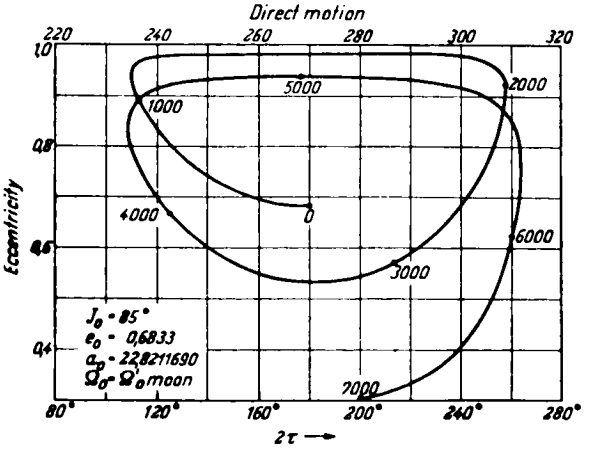


FIGURE 7.

long range variations of elements of comets and minor planets. The variations of the elements of Encke's comet were computed as an example (fig. 9 and 10). The results agree basically with the results obtained by Whipple and Hamid (1950), but the Halphen–Goriachev method gives a period which is somewhat longer than that given by Whipple.

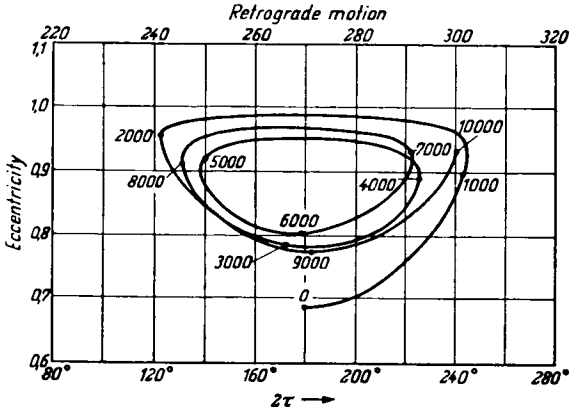


FIGURE 8.

Let M_0, ω, Ω, i, e, a, n, be the osculating elements of the artificial satellite referred to the Earth's equator and let M_0', ω', Ω', i', e', a', u', m' be the elements and the mass of the moon referred to a fixed ecliptic and equinox. Put

$$A_1(\alpha)=\begin{bmatrix} +1 & 0 & 0 \\ 0 & +\cos\alpha & -\sin\alpha \\ 0 & +\sin\alpha & +\cos\alpha \end{bmatrix}$$

$$A_3(\alpha)=\begin{bmatrix} +\cos\alpha & -\sin\alpha & 0 \\ +\sin\alpha & +\cos\alpha & 0 \\ 0 & 0 & +1 \end{bmatrix}.$$

The following system of formulae, combined from the system given by Goriachev (1937) and from the system given in author's previous work (1961) was used for the actual computations of the long range lunar effects:

$$[\mathbf{P}, \mathbf{Q}, \mathbf{R}]=A_3(\Omega)\cdot A_1(i)\cdot A_3(\omega)$$

$$[\mathbf{P}', \mathbf{Q}', \mathbf{R}']=A_1(\epsilon)\cdot A_3(\Omega')\cdot A_1(i')\cdot A_3(\omega')$$

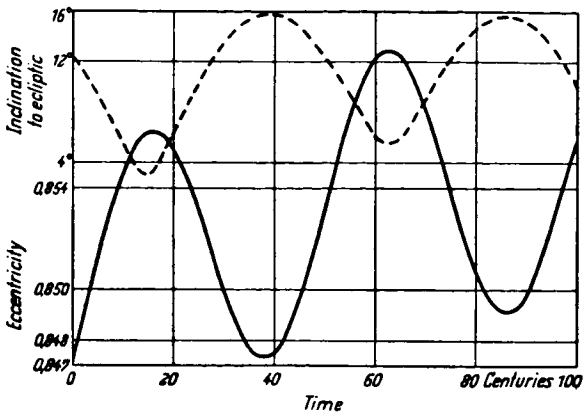


FIGURE 9.—Secular perturbations of Encke's comet.

$$\mathbf{r}=\mathbf{P}a(\cos u-e)+\mathbf{Q}a(1-e^2)^{1/2}\sin u$$

$$\varrho=\mathbf{P}(a/a')(\cos u-e)+\mathbf{Q}(a/a')\times(1-e^2)^{1/2}\sin u+e'\mathbf{P}'$$

$$\alpha=\varrho\cdot\mathbf{P}', \quad \beta=\varrho\cdot\mathbf{Q}', \quad \gamma=\varrho\cdot\mathbf{R}'$$

$$K_1=\rho^2-2+e'^2$$

$$K_2=(1-e'^2)(1-\alpha^2)-\beta^2-(2-e'^2)\gamma^2$$

$$K_3=\gamma^2(1-e'^2)$$

$$g_2=\tfrac{4}{3}(K_1^2-3K_2)$$

$$g_3=\tfrac{4}{27}(2K_1^3-9K_1K_2+27K_3)$$

$$\xi=27g_3^2/g_2^3$$

$$A=\frac{6^{1/2}(g_2)^{1/4}}{9g_2^3}\,\frac{144}{\pi}\,(\xi)^{1/2}\psi'(\xi)$$

$$B=\frac{\sqrt{2}}{\pi g_2(g_2)^{1/4}}\psi(\xi)$$

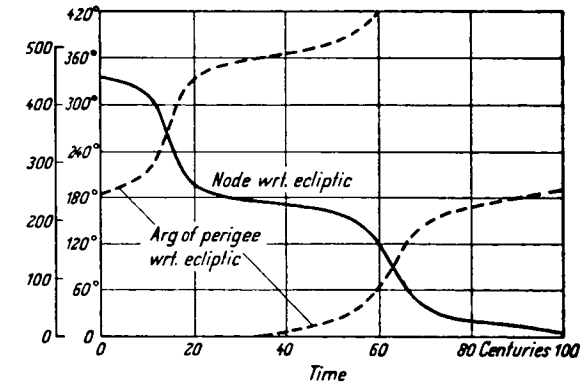


FIGURE 10.—Secular perturbations of Encke's comet.

$$w=\frac{1-(\xi)^{1/2}}{1+(\xi)^{1/2}}$$

$$\psi(\xi)=\left(\frac{2}{1+(\xi)^{1/2}}\right)^{1/6}$$

$$\cdot\frac{144}{\pi}(\xi)^{1/2}\psi'(\xi)=\left(\frac{2}{1+(\xi)^{1/2}}\right)^{7/6}.$$

$$\begin{aligned}
&(+2.3870942 && (-3.7991784\\
&-0.0663082 && +0.3693646w\\
&+0.0225632w^2 && -0.1556119w^2\\
&-0.0117691w^3 && +0.0889726w^3\\
&+0.0073743w^4 && -0.0586828w^4\\
&-0.0051060w^5 && +0.0419870w^5\\
&+0.0037250w^6 && -0.0313364w^6\\
&-0.0027325w^7 && +0.0233758w^7\\
&+0.0019070w^8 && -0.0165247w^8\\
&-0.0011936w^9 && +0.0104483w^9\\
&+0.0006337w^{10} && -0.0055933w^{10}\\
&-0.0002710w^{11} && +0.0024083w^{11}\\
&+0.0000884w^{12} && -0.0007898w^{12}\\
&-0.0000205w^{13} && +0.0001837w^{13}\\
&+0.0000030w^{14} && -0.0000268w^{14}\\
&-0.0000002w^{15}) && +0.0000018w^{15})
\end{aligned}$$

$$K_4=9K_3-K_1K_2$$

$$K_5=K_1(K_1K_2-3K_3)-2K_2^2$$

$$a_{11}=K_4(\alpha^2-1)+K_5+\tfrac{3}{2}g_2K_3$$

$$a_{22}=K_4(\beta^2-1+e'^2)+K_5+\tfrac{3}{2}[g_2K_3/(1-e'^2)]$$

$$a_{33}=K_4\gamma^2+K_5+\tfrac{3}{2}g_2[\alpha^2(1-e'^2)+\beta^2-(1-e'^2)]$$

$$a_{12}=a_{21}=K_4\alpha\beta$$

$$a_{23}=a_{32}=(K_4-\tfrac{3}{2}g_2)\beta\gamma$$

$$a_{31}=a_{13}=[K_4-\tfrac{3}{2}g_2(1-e'^2)]\gamma\alpha$$

$$a'_{11}=\alpha^2-1-\tfrac{1}{3}K_1$$

$$a'_{22}=\beta^2-1+e'^2-\tfrac{1}{3}K_1$$

$$a'_{33}=\gamma^2-\tfrac{1}{3}K_1$$

$$a'_{12}=a'_{21}=\alpha\beta$$

$$a'_{23}=a'_{32}=\beta\gamma$$

$$a'_{31}=a'_{13}=\gamma\alpha$$

Then the dyadic Φ is formed

$$\begin{aligned}
\mathbf{\Phi}=&+A_{11}\mathbf{P}'\mathbf{P}'+A_{12}\mathbf{P}'\mathbf{Q}'+A_{13}\mathbf{P}'\mathbf{R}'\\
&+A_{21}\mathbf{Q}'\mathbf{P}'+A_{22}\mathbf{Q}'\mathbf{Q}'+A_{23}\mathbf{Q}'\mathbf{R}'\\
&+A_{31}\mathbf{R}'\mathbf{P}'+A_{32}\mathbf{R}'\mathbf{Q}'+A_{33}\mathbf{R}'\mathbf{R}'
\end{aligned}$$

$$A_{ij}=a_{ij}A+a_{ij}B$$

The disturbing "force" averaged over the orbit of the moon is

$$F_0=-(2km'/a'^3)\,\mathbf{\Phi}\mathbf{r}.$$

where k is the Gaussian constant.

S_0, T_0, W_0, the components of the disturbing force in the direction of $\mathbf{r}$, $\mathbf{R}\times\mathbf{r}$ and $\mathbf{R}$, are

$$S_0=-(2km'/a'^3)r^{-1}\mathbf{r}\cdot\mathbf{\Phi}\cdot\mathbf{r}$$

$$T_0=-(2km'/a'^3)r^{-1}(\mathbf{R}\times\mathbf{r})\cdot\mathbf{\Phi}\cdot\mathbf{r}$$

$$W_0=-(2km'/a'^3)\,\mathbf{R}\cdot\mathbf{\Phi}\cdot\mathbf{r}.$$

The equations for the variation of elements take the form:

$$\frac{de}{dt}=\frac{[a(1-e^2)]^{1/2}}{2\pi}\int_0^{2\pi}[(1-e_2)^{1/2}S_0\sin u+T_0\times(-\tfrac{3}{2}e+2\cos u-\tfrac{1}{2}e\cos 2u)]\,du$$

$$\frac{di}{dt}=\left(\frac{a}{1-e^2}\right)^{1/2}\frac{1}{2\pi}\int_0^{2\pi}W_0\frac{r}{a}[(\cos u-e)\times\cos\omega-(1-e^2)^{1/2}\sin u\sin\omega]\,du$$

$$\sin i\frac{d\Omega}{dt}=\left(\frac{a}{1-e^2}\right)^{1/2}\frac{1}{2\pi}\int_0^{2\pi}W_0\frac{r}{a}[(\cos u-e)\times\sin\omega+(1-e^2)^{1/2}\sin u\cos\omega]\,du$$

$$\frac{d\pi}{dt}=\frac{[a(1-e^2)]^{1/2}}{e}\frac{1}{2\pi}\int_0^{2\pi}\left[-S_0(\cos u-e)+T_0\times\left(1+\frac{r}{a}\frac{1}{1-e^2}\right)(1-e^2)^{1/2}\sin u\right]du,$$

where $\pi=\omega+\Omega$.

The solar perturbations present less theoretical and computational difficulties, than do the lunar perturbations. The system of formulae (45′), (46′), (47′), (57) derived in the author's previous work (1961) is adequate to take all significant long period terms into account. This system is in use in the Theoretical Division for the determina-

tion of the lifetime and the long range stability of artificial satellites.

Tesseral harmonics in the earth's potential can also introduce some long range effects, especially if the mean motion of the satellite is commensurable with the angular velocity of earth's rotation. The most interesting case of long range effects produced by tesseral harmonics is the influence of the ellipticity of earth's equator on the motion of a 24-hour satellite.

This effect was treated by L. Blitzer, E. M. Boughton, G. Kang and R. M. Page (1962) using a linearized system of differential equations of motion of the satellite and by Morando (1962) using von Zeipel's method. Sehnal (1960) treated the same problem under the condition that the commensurability is not very sharp and that it would be possible to develop the integrals of the problem into trigonometric series. He made use of the elements of Delaunay

$$L,\ G,\ H,\ l,\ g,\ h.$$

The author and A. Bailie (1962) treated the problem using the canonincal elements

$$x_1=L, \qquad y_1=l+g+h-n't$$
$$x_2=L-G, \qquad y_2=-g$$
$$x_3=L-H, \qquad y_3=-h+n't$$

The disturbing function F was taken in the form

$$F=F_0+Q_0\cos 2y_1+Q_1\cos(2y_1+2y_2)+Q_2\cos(2y_1+4y_2),$$

where F_0 is Brouwer's expression (1959) for the secular part of the disturbing function to which the term $n'H$ is added. Brouwer's expression is transformed to the set x_1, x_2, x_3. The coefficients Q_0, Q_1, Q_2 are:

$$Q_0=+\tfrac{3}{8}(\mu A_{22}/a^3)(1+\cos i)^2\times(1-\tfrac{5}{2}e^2+\tfrac{13}{16}e^4-\tfrac{35}{288}e^6)$$

$$Q_1=+\tfrac{3}{4}(\mu A_{22}/a^3)\sin^2 i(\tfrac{9}{4}e^2+\tfrac{7}{4}e^4+\tfrac{141}{64}e^6)$$

$$Q_2=+\tfrac{3}{8}(\mu A_{22}/a^3)(1-\cos i)^2(\tfrac{1}{24}e^4+\tfrac{7}{240}e^6),$$

where μ is the gravitational constant and A_{22} is the geodetic parameter associated with the Ω_2^2 tesseral harmonic.

Two cases are possible:

1) A stable case, when the conditions

$$(\partial F_0/\partial x_1)-(\partial Q_0/\partial x_1)=0$$

for

$$y_1=\tfrac{1}{2}\pi,\ \tfrac{3}{2}\pi$$

are satisfied and

2) An un-stable where

$$(\partial F_0/\partial x_1)+(\partial Q_0/\partial x_1)=0$$

for

$$y_1=0,\ 2\pi.$$

The problem admits two integrals, the integral of energy

$$F=-C$$

and

$$x_3=\alpha_3.$$

Let α_1, α_2, α_3 be the constants of integration associated with x_1, x_2, x_3 and for the stable case put

$$R_0=F_0-Q_0$$

and for the unstable case put

$$R_0=F_0+Q_0$$

and let

$$w_{ij}=\frac{\partial^{i+j}}{\partial\alpha_1{}^i\partial\alpha_2{}^j}R_0(\alpha_1,\ \alpha_2,\ \alpha_3).$$

The case when w_{10} is small is a near resonance case and the integrals are not developable into trigonometric series. Bohlin's resonance theory (1889) can be used to solve the problem. The complete integral of the Hamiltonian partial differential equation

$$F=-C_0-C_1-\cdots,$$

with

$$x_1=\partial S/\partial y_1, \qquad x_2=\partial S/\partial y_2,$$

is developed into series of the form

$$S=\alpha_1y_1+\alpha_2y_2+\alpha_3y_3+\varphi(y_1)+[\varphi_2(y_1)+\psi_2(y_1,\ y_2)]+\cdots$$

with respect to the parameter $w=w_{10}/w_{20}$. The function ψ_2 is a purely trigonometric function. In the exposition, the value of $Q_0(\alpha_1,\ \alpha_2,\ \alpha_3)$ was considered to be of the order of w^2.

The constant of energy is also decomposed into

a series in w. This decomposition served to remove possible zero divisors from the higher approximations by the particular choices of C_1, C_2, $\cdots$. The constant C_0 is defined as

$$C_0 = -R_0(\alpha_1, \alpha_2, \alpha_3).$$

For the stable case the solution of the Hamiltonian was obtained in the form

$$S = (\alpha_1 + A_1)y_1 + \alpha_2 y_2 + \alpha_3 y_3 \pm \int (A_2 - A_3 \cos^2 y_1)^{1/2}\, dy_1 + A_4 \sin 2y_1 + A_5 \sin(2y_1 + 2y_2) + A_6 \sin(2y_1 + 4y_2) + \cdots$$

and for the un-stable case in the form

$$S = (\alpha_1 + B_1)y_1 + \alpha_2 y_2 + \alpha_3 y_3 + \int (B_2 + B_3 \sin^2 y_1)^{1/2}\, dy_1 + B_4 \sin 2y_1 + B_5 \sin(2y_1 + 2y_2) + B_6 \sin(2y_1 + 4y_2) + \cdots$$

The coefficients A_i and B_i are power series in w with the coefficients depending upon α_1, α_2, α_3. They were obtained by the authors with accuracy up to w^2. Determination of the integrals of motion in the form

$$x_i = \partial S/\partial y_i \qquad i = 1, 2, 3$$
$$(\partial C/\partial \alpha_i)t = \beta_i = \partial S/\partial \alpha_i$$

did not present any difficulty.

For orbits with small eccentricities the terms having the arguments $2y_1+2y_2$, $2y_1+4y_2$ are very small long period terms. Thus, the main part of the motion of the point $(x_1, x_2, x_3; y_1, y_2, y_3)$ in its phase space in a stable case consists of describing ovals around the libration points in the plane (x_1, y_1). The period of this motion is

$$T = \frac{\pi}{(Q_0 w_{20})^{1/2}}\left\{1 - w\left[\frac{2}{3\pi}\frac{w_{30}}{w_{20}} + \frac{Q_0'}{Q_0} - \frac{1}{2w_{20}} \times \left(\frac{w_{30}}{w_{20}} + 3\frac{Q_0'}{Q_0}\right)\right]\right\} + O(w^2).$$

In the unstable case the argument y_1 possesses a real secular term, which is not present in the stable case. Its value is

$$\nu_1 = n_1/P,$$

where

$$n_1 = -w w_{20} + \tfrac{1}{2}\left(\frac{w_{30}}{w_{20}} + 3\frac{Q_0'}{Q_0}\right)w^2$$

$$P = -\tfrac{1}{3}w\frac{w_{30}}{w_{20}} + \frac{2k'K}{\pi}\left(1 - \frac{Q_0'}{Q_0}\right),$$

$$Q_0' = (\partial/\partial\alpha_1)Q_0(\alpha_1, \alpha_2, \alpha_3),$$

K is a complete elliptic integral of the first kind with the modulus

$$k = [B_3/(B_2 + B_3)]^{1/2}$$

and the complementary modulus

$$k' = (1 - k^2)^{1/2}.$$

REFERENCES

1. Blitzer, L., Boughton, E. M., Kang, G., and Page, R. M.: Effect of ellipticity of the equator on 24 hour nearly circular satellite orbits, J. G. R. **67,** 329 (1962).
2. Bohlin, K. P., Uber eine neue Annäherungsmethode in der Störungstheorie, Ak. Handl. Bihang, **14** (Afd. 1). Stockholm (1889).
3. Brouwer, D., Solution of the problem of artificial satellite theory without drag, Astronom. J., **64,** 378 (1959).
4. Brown, E. W., and Shook, C. A., Planetary Theory. Cambridge, 159 (1933).
5. Gauss, K. F., Determinatio attractionis quam in punctum quodvis positionis data exerceret planeta, si eius massa, etc., Coll. Works, III, 331 (1818).
6. Goriachev, N. N., On the method of Halphen of the computation of secular perturbations (Russ.), University of Tomsk., 1–115 (1937).
7. Goursat, E., Ann. Scient. École Norm. Sup. (2), **10,** 3–142 (1881).
8. Halphen, G. H., Traite des fonction élliptiques, II, Paris (1888).
9. Hill, G. W., Secular perturbations of the planets, Amer. J. Math. **23,** 317–336 (1901).
10. Hirayama, K., Proc. Imp. Acad. Jap., **3,** 9 (1927).
11. Kozai, Y., Private communication (1959).
12. Morando, B., Libration d'un satellite de 24 h., C. R. Acad. Sci., Paris du 22 Jan., 635 (1962).
13. Musen, P., On the long period lunisolar effect in the motion of the artificial satellite, J. G. R., **66,** 1659 (1961).
14. Musen, P., and Bailie, A., On the motion of a 24-hour satellite, J. G. R., **67,** 1123 (1962).
15. Musen, P., Bailie, A., and Upton, E., Development of the lunar and solar perturbations in the motion of an artificial satellite. NASA TN D-494, January (1961).
16. Sehnal, L., The influence of the equatorial ellipticity of the earth gravitational field on the motion of a close satellite, Bull. Astron. Inst. Czechoslovakia, **11** (3), 90–93 (1960).
17. Whipple, F. L., and Hamid, S. E. D., Harv. Abstract, **9,** 248 (1950).

ON A MODIFICATION OF HANSEN'S LUNAR THEORY

PETER MUSEN

Goddard Space Flight Center

A modification of Hansen's lunar theory is given in a form that permits a purely numerical treatment of solar perturbations of planetary satellites or of perturbations caused by the earth in the motion of hypothetical lunar satellites. The development of the coordinates and of the velocities of a satellite is obtained in the form of a trigonometric series in four arguments with numerical coefficients. At the earlier stages of the computation we are not compelled to use Hansen's basic W-function: We make use of the fictitious mean anomaly and the replacement "bar" operation—two characteristic features of Hansen's theory—to set up a process of iteration in a convenient form, rather than to compute W. The preference given to the method of iteration was motivated by the circumstance that input information must not necessarily be limited to the choice of the rotating ellipse as an intermediary orbit. A systematic investigation of the orbital stability of lunar satellites can be based on the development given herein.

INTRODUCTION

In this paper a modification of Hansen's lunar theory[1] is given in a form that permits a purely numerical treatment of solar perturbations of planetary satellites as well as the perturbations caused by the earth in the motion of hypothetical lunar satellites. The development of perturbations of very close lunar satellites does not present any difficulty and therefore is not given here. The case case of more distant satellites was treated by Lemekhova[2] on the basis of Delaunay's theory for orbits with low inclinations toward the orbit of the moon. At the present time we are not in possession of a complete analytical theory of satellites with highly inclined orbits, although Delaunay's theory can be modified to become valid for such orbits. Because of this circumstance, the use of a semianalytical theory is suggested, as given in this paper.

Hansen's original lunar theory requires the use of the derivatives of the disturbing functions with respect to the eccentricity and the parameters determining the orbit plane's position. This means that a literal development of the disturbing function must be obtained before the numerical values of the elements can be substituted. However, with the advent of electronic machines, more emphasis can be put on the development of purely numerical theories.

Unlike Hansen's lunar theory, his planetary theory[3] is a purely numerical one. For this reason the theory of perturbations in the orbit plane, as presented herein, takes Hansen's planetary theory, and not his lunar theory, as a pattern. In addition, at the earlier stages of the computations we are not compelled to use Hansen's basic W-function. Instead, we make use of the fictitious mean anomaly and the replacement "bar" operation to set up the process of iteration in a convenient form, rather than to compute W. The preference given to the method of iteration over the method of computing perturbations of different orders was motivated by the circumstance that input information must not necessarily be limited to the choice of Hansen's rotating ellipse as an intermediary orbit. In addition, the programing and the computational process become more uniform if the method of iteration is used.

Only after the process of iteration is completed

is the $\bar{W}$-function formed in order to determine the perturbations in the mean anomaly and in the radius vector.

The orbit of the disturbing body is taken to be a moving ellipse.

The inclination of the satellite's orbit plane toward the orbit plane of the disturbing body can be large. For this reason we do not deduce the standard development of the latitude into series, but make use of four "redundant" parameters introduced by the author in the theory of artificial satellites.[4] The introduction of these parameters makes all basic arguments linear in time from the outset, and it permits us to deduce the rectangular components of the position and velocity vectors very easily.

The knowledge of these components facilitates the computation of osculating elements.

To simplify the form of the development, we take—as did Hansen—the orbit plane of the disturbing body as a basic reference plane.

The effect of this plane's motion on the relative motion of the satellite is very small and can easily be taken into consideration.

The form of the disturbing function in our case favors only the choice of time as an independent variable, but not the true or the eccentric anomaly as in the case of an artificial satellite of the earth.

DEVELOPMENT OF THE DISTURBING FUNCTION

For a satellite of negligible mass, we have

$$\Omega = \frac{m' r^2}{r'^3}\left(\frac{3}{2}S^2 - \frac{1}{2}\right) + \frac{m' r^3}{r'^4}\left(\frac{5}{2}S^3 - \frac{3}{2}S\right) + \frac{m' r^4}{r'^5}\left(\frac{35}{8}S^4 - \frac{15}{4}S^2 + \frac{3}{8}\right) + \cdots . \quad (1)$$

(See app. A for list of notations.) Putting

$$s = \frac{r}{a}\cdot\frac{a'}{r'}S\ , \qquad p = \frac{r}{a}\cdot\frac{a'}{r'}\ , \quad (2)$$

we deduce:

$$\Omega = \frac{m' a^2}{a'^3}\cdot\frac{a'}{r'}\cdot\left(\frac{3}{2}s^2 - \frac{1}{2}p^2\right) + \frac{m' a^3}{a'^4}\cdot\frac{a'}{r'}\cdot\left(\frac{5}{2}s^3 - \frac{3}{2}p^2 s\right) + \frac{m' a^4}{a'^5}\cdot\frac{a'}{r'}\left(\frac{35}{8}s^4 - \frac{15}{4}s^2 p^2 + \frac{3}{8}p^4\right) + \cdots , \quad (3)$$

$$r\frac{\partial\Omega}{\partial r} = a\frac{\partial\Omega}{\partial a} = 2\frac{m' a^2}{a'^3}\cdot\frac{a'}{r'}\cdot\left(\frac{3}{2}s^2 - \frac{1}{2}p^2\right) + 3\frac{m' a^3}{a'^4}\cdot\frac{a'}{r'}\left(\frac{5}{2}s^3 - \frac{3}{2}p^2 s\right) + 4\frac{m' a^4}{a'^5}\cdot\frac{a'}{r'}\cdot\left(\frac{35}{8}s^4 - \frac{15}{4}s^2 p^2 + \frac{3}{8}p^4\right) + \cdots , \quad (4)$$

and

$$\frac{a}{r}\cdot\frac{r'}{a'^2}\frac{\partial\Omega}{\partial S} = 3\frac{m'a^2}{a'^3}\cdot\frac{a'}{r'}s + \frac{m'a^3}{a'^4}\cdot\frac{a'}{r'}\left(\frac{15}{2}s^2-\frac{3}{2}p^2\right) + \frac{m'a^4}{a'^5}\cdot\frac{a'}{r'}\left(\frac{35}{2}s^3-\frac{15}{2}p^2s\right) . \tag{5}$$

We have

$$S = \cos(v-\psi)\cos(v'-\psi') + \sin(v-\psi)\sin(v'-\psi')\cos I , \tag{6}$$

$$s = \frac{r}{a}\cdot\frac{a'}{r'}\cos(v-\psi)\cos(v'-\psi') + \frac{r}{a}\cdot\frac{a'}{r'}\sin(v-\psi)\sin(v'-\psi')\cos I . \tag{6'}$$

Taking the relations

$$2N = \psi_0 + \psi_0' - \psi - \psi' - 2n_0\alpha t ,$$

$$2K = \psi_0 - \psi_0' - \psi + \psi' + 2n_0\eta t ,$$

$$v = \bar{f} + \pi_0 + n_0 y t ,$$

$$v' = \bar{f}' + \pi_0' + n_0 y' t$$

into account, we deduce from the equations 6 and 6′ that

$$S = +(\lambda_1^2-\lambda_2^2)\cos(\bar{f}+\bar{f}'+\omega+\omega') - 2\lambda_1\lambda_2\sin(\bar{f}+\bar{f}'+\omega+\omega') + (\lambda_4^2-\lambda_3^2)\cos(\bar{f}-\bar{f}'+\omega-\omega') - 2\lambda_3\lambda_4\sin(\bar{f}-\bar{f}'+\omega-\omega') , \tag{7}$$

$$s = (1+\nu)\left[+(\lambda_1^2-\lambda_2^2)\frac{\bar{r}}{a}\cdot\frac{a'}{\bar{r}'}\cos(\bar{f}+\bar{f}'+\omega+\omega') - 2\lambda_1\lambda_2\frac{\bar{r}}{a}\cdot\frac{a'}{\bar{r}'}\sin(\bar{f}+\bar{f}'+\omega+\omega') + (\lambda_4^2-\lambda_3^2)\frac{\bar{r}}{a}\cdot\frac{a'}{\bar{r}'}\cos(\bar{f}-\bar{f}'+\omega-\omega') - 2\lambda_3\lambda_4\frac{\bar{r}}{a}\cdot\frac{a'}{\bar{r}'}\sin(\bar{f}-\bar{f}'+\omega-\omega')\right] , \tag{7'}$$

where

$$\omega = \omega_0 + n_0 \left(y + \alpha - \eta\right) t \ , \tag{8}$$

$$\omega_0 = \pi_0 - \psi_0 \ ,$$

$$\omega' = \omega_0' + n_0 \left(y' + \alpha + \eta\right) t \ ,$$

$$\omega_0' = \pi_0' - \psi_0' \ . \tag{9}$$

The arguments ω and ω' are defined as linear functions of time from the very start.

The next step is to obtain the developments of

$$\frac{\bar{r}}{a} \cdot \frac{a'}{\bar{r}'} \begin{matrix} \cos \\ \sin \end{matrix} \left(\bar{f} \pm \bar{f}' + \omega \pm \omega'\right)$$

and

$$\frac{\bar{r}}{a} \cdot \frac{a'}{\bar{r}'} \ .$$

These five series represent an input. They can be easily computed, even using a desk calculator, by means of Cayley's scheme as shown on p. 415.

In this scheme we multiply every line of the upper part by every line of the lower part and add the arguments together, then attach the sin or cos symbol to each term of the product. The upper sign of the argument in the solar part is associated with the development of

$$\frac{\bar{r}}{a} \cdot \frac{a'}{\bar{r}'} \begin{matrix} \cos \\ \sin \end{matrix} \left(\bar{f} + \bar{f}' + \omega + \omega'\right) \ ,$$

and the lower sign is associated with the development of

$$\frac{\bar{r}}{a} \cdot \frac{a'}{\bar{r}'} \begin{matrix} \cos \\ \sin \end{matrix} \left(\bar{f} - \bar{f}' + \omega - \omega'\right) \ .$$

The numerical values of e and e' can be substituted from the very start.

$\frac{\bar{r}}{a} \cdot \frac{a'}{\bar{r}'} \begin{matrix}\cos \\ \sin\end{matrix} (\bar{f} \pm \bar{f}' + \omega \pm \omega')$	sin or cos	
. . . . .		
$-\frac{1}{12} e'^3 - \frac{1}{24} e'^5 - \frac{7}{240} e'^7 + \cdots$	$\mp$ 2g'	
$-\frac{1}{8} e'^2 - \frac{1}{16} e'^4 - \frac{115}{3072} e'^6 + \cdots$	$\mp$ g'	
$-\frac{1}{2} e' - \frac{1}{8} e'^3 - \frac{1}{16} e'^5 - \frac{5}{128} e'^7 + \cdots$	0.g'	$\pm \omega'$
$+ 1 - e'^2 - \frac{1}{64} e'^4 - \frac{17}{288} e'^6 + \cdots$	$\pm$ g'	
$+\frac{3}{2} e' - \frac{7}{4} e'^3 + \frac{31}{96} e'^5 - \frac{1}{12} e'^7 + \cdots$	$\pm$ 2g'	
$+\frac{17}{8} e'^2 - \frac{47}{16} e'^4 + \frac{1069}{1024} e'^6 - \cdots$	$\pm$ 3g'	
. . . . .	. . . .	
. . . . .	. . . .	
$+\frac{125}{9216} e^6 + \cdots$	- 5g	
$+\frac{1}{60} e^5 - \frac{1}{720} e^7 + \cdots$	- 4g	
$+\frac{3}{128} e^4 + \frac{3}{1280} e^6 + \cdots$	- 3g	
$+\frac{1}{24} e^3 + \frac{1}{96} e^5 + \frac{1}{120} e^7 + \cdots$	- 2g	
$+\frac{1}{8} e^2 + \frac{1}{24} e^4 + \frac{25}{1024} e^6 + \cdots$	- g	
$-\frac{3}{2} e$ (exact)	0.g	$+ \omega$
$+ 1 - \frac{1}{2} e^2 - \frac{1}{64} e^4 - \frac{29}{1152} e^6 + \cdots$	+ g	
$+\frac{1}{2} e - \frac{3}{8} e^3 + \frac{5}{96} e^5 - \frac{1}{72} e^7 + \cdots$	+ 2g	
$+\frac{3}{8} e^2 - \frac{3}{8} e^4 + \frac{111}{1024} e^6 + \cdots$	+ 3g	
$+\frac{1}{3} e^3 - \frac{5}{12} e^5 + \frac{43}{240} e^7 + \cdots$	+ 4g	
$+\frac{125}{384} e^4 - \frac{125}{256} e^6 + \cdots$	+ 5g	
. . . . .	. . . .	

In the same manner we have:

$\frac{\bar{r}}{a} \cdot \frac{a'}{\bar{r}'}$	cos
. . . . .	
$+ \frac{9}{16} e'^3 - \frac{81}{256} e'^5 + \frac{729}{10240} e'^7 - \cdots$	$- 3g'$
$+ \frac{1}{2} e'^2 - \frac{1}{6} e'^4 + \frac{1}{48} e'^6 - \cdots$	$- 2g'$
$+ \frac{1}{2} e' - \frac{1}{16} e'^3 + \frac{1}{384} e'^5 - \frac{1}{18432} e'^7$	$- g'$
$+ 1$	$0.g'$
$+ \frac{1}{2} e' - \frac{1}{16} e'^3 + \frac{1}{384} e'^5 - \frac{1}{18432} e'^7$	$+ g'$
$+ \frac{1}{2} e'^2 - \frac{1}{6} e'^4 + \frac{1}{48} e'^6 - \cdots$	$+ 2g'$
$+ \frac{9}{16} e'^3 - \frac{81}{256} e'^5 + \frac{729}{10240} e'^7 - \cdots$	$+ 3g'$
. . . . .	. . . .
. . . . .	. . . .
$- \frac{125}{768} e^5 + \cdots$	$- 5g$
$- \frac{1}{6} e^4 + \frac{1}{5} e^6 + \cdots$	$- 4g$
$- \frac{3}{16} e^3 + \frac{45}{256} e^5 - \frac{567}{10240} e^7 + \cdots$	$- 3g$
$- \frac{1}{4} e^2 + \frac{1}{6} e^4 - \frac{1}{32} e^6 + \cdots$	$- 2g$
$- \frac{1}{2} e + \frac{3}{16} e^3 - \frac{5}{384} e^5 + \frac{7}{18432} e^7 - \cdots$	$- g$
$\frac{\bar{r}}{a} \cdot \frac{a'}{\bar{r}'}$	cos
$+ 1 + \frac{1}{2} e^2$	$0.g$
$- \frac{1}{2} e + \frac{3}{16} e^3 - \frac{5}{384} e^5 + \frac{7}{18432} e^7 - \cdot$	$+ g$
$- \frac{1}{4} e^2 + \frac{1}{6} e^4 - \frac{1}{32} e^6 + \cdots$	$+ 2g$
$- \frac{3}{16} e^3 + \frac{45}{256} e^5 - \frac{567}{10240} e^7 + \cdots$	$+ 3g$
$- \frac{1}{6} e^4 + \frac{1}{5} e^6 + \cdots$	$+ 4g$
$- \frac{125}{768} e^5 + \cdots$	$+ 5g$
. . . . .	. . . .

PERTURBATIONS IN THE ORBIT PLANE

To determine the perturbations in the radius vector and in the mean anomaly, Hansen from the very start introduces the function

$$W = \Xi + \Upsilon\left(\frac{\rho}{a_0}\cos\varphi + \frac{3}{2}e_0\right) + \Psi \cdot \frac{\rho}{a_0}\sin\varphi \ , \tag{10}$$

where

$$\Xi = -1 - \frac{h_0}{h} + 2\frac{h}{h_0} - 3e_0\frac{h}{h_0} \cdot \frac{e\cos\left(\chi - n_0 yt - \pi_0\right) - e_0}{1 - e_0^2} \ , \tag{11}$$

$$\Upsilon = 2\frac{h}{h_0} \cdot \frac{e\cos\left(\chi - n_0 yt - \pi_0\right) - e_0}{1 - e_0^2} \ , \tag{12}$$

$$\Psi = 2\frac{h}{h_0} \cdot \frac{e\sin\left(\chi - n_0 yt - \pi_0\right)}{1 - e_0^2} \ . \tag{13}$$

From the standpoint of the process of iteration the use of the elements

$$\frac{h_0}{h}, \quad \Upsilon, \quad \Psi$$

would be preferable. We have[1]

$$\frac{d}{dt}\frac{h_0}{h} = h_0\frac{\partial\Omega}{\partial f} \ , \tag{14}$$

$$\frac{d\Upsilon}{dt} = n_0 y\Psi + \frac{2a_0 n_0}{\sqrt{1 - e_0^2}}\left\{\left[\left(\frac{a_0}{r} + \frac{h^2}{h_0^2}\cdot\frac{1}{1 - e_0^2}\right)\cos\bar{f} + \frac{e_0}{1 - e_0^2}\frac{h^2}{h_0^2}\right]\frac{\partial\Omega}{\partial f} + \frac{a_0}{r}\sin\bar{f}\left(r\frac{\partial\Omega}{\partial r}\right)\right\} \ , \tag{15}$$

$$\frac{d\Psi}{dt} = -n_0 y\left(\Upsilon + 2\frac{h}{h_0}\cdot\frac{e_0}{1 - e_0^2}\right) + \frac{2a_0 n_0}{\sqrt{1 - e_0^2}}\left\{\left(\frac{a_0}{r} + \frac{h^2}{h_0^2}\cdot\frac{1}{1 - e_0^2}\right)\sin\bar{f}\frac{\partial\Omega}{\partial f} - \frac{a_0}{r}\cos\bar{f}\cdot\left(r\frac{\partial\Omega}{\partial r}\right)\right\} . \tag{16}$$

Taking into account

$$\frac{\partial\Omega}{\partial f} = \frac{1}{\sqrt{1 - e_0^2}}\left(\frac{\bar{r}^2}{a_0^2}\frac{\partial\Omega}{\partial l} - \frac{\bar{r}}{a_0}\cdot\frac{e_0\sin\bar{f}}{\sqrt{1 - e_0^2}}\cdot r\frac{\partial\Omega}{\partial r}\right) , \tag{17}$$

we transform equation 15 to the form

$$\frac{d\Upsilon}{dt} = +n_0 y\Psi + M_1\frac{\partial\Omega}{\partial l} + N_1 \cdot r\frac{\partial\Omega}{\partial r} \ , \tag{18}$$

where

$$M_1 = \frac{2a_0 n_0}{1 - e_0^2}\left[\frac{1}{e_0}\left(1 - e_0^2 - \frac{\bar{r}^2}{a_0^2}\right) - \frac{\nu}{1+\nu}\frac{1}{e_0}\left(1 - e_0^2 - \frac{\bar{r}}{a_0}\right) + \left(\frac{h^2}{h_0^2} - 1\right)\frac{1}{e_0}\frac{\bar{r}}{a_0}\left(1 - \frac{\bar{r}}{a_0}\right)\right] \quad (19)$$

and

$$N_1 = \frac{2a_0 n_0}{1 - e_0^2} \cdot \frac{\bar{r}}{a_0}\frac{\sin \bar{f}}{\sqrt{1 - e_0^2}}\left[1 - \frac{a_0}{\bar{r}} \cdot \frac{\nu}{1+\nu} - \left(\frac{h^2}{h_0^2} - 1\right)\left(\frac{a_0}{\bar{r}} - 1\right)\right] . \quad (20)$$

In a similar manner, using the basic relations

$$\frac{\bar{r}}{a_0}\sin \bar{f} + \frac{\bar{r}^2}{a_0^2}\frac{\sin \bar{f}}{1 - e_0^2} = \int \left(2\frac{\bar{r}}{a_0}\cos \bar{f} + 3e_0\right) dl ,$$

$$\frac{d}{dl}\frac{\bar{r}}{a_0}\cos \bar{f} = -\frac{\sin \bar{f}}{\sqrt{1 - e_0^2}} ,$$

$$\frac{d}{dl}\frac{\bar{r}}{a_0}\sin \bar{f} = \frac{\cos \bar{f} + e_0}{\sqrt{1 - e_0^2}} ,$$

we deduce from equation 16:

$$\frac{d\Psi}{dt} = -n_0 y\left(\Upsilon + 2\frac{h}{h_0}\frac{e_0}{1 - e_0^2}\right) + M_2\frac{\partial \Omega}{\partial l} + N_2 \cdot r\frac{\partial \Omega}{\partial r} , \quad (21)$$

where

$$M_2 = \frac{2a_0 n_0}{1 - e_0^2}\left[\frac{1}{\sqrt{1 - e_0^2}}\int \left(2\frac{\bar{r}}{a_0}\cos \bar{f} + 3e_0\right) dl - \frac{\nu}{1+\nu} \cdot \frac{\bar{r}}{a_0}\sin \bar{f} + \left(\frac{h^2}{h_0^2} - 1\right) \cdot \frac{\bar{r}^2}{a_0^2} \cdot \frac{\sin \bar{f}}{1 - e_0^2}\right] , \quad (22)$$

$$N_2 = \frac{2a_0 n_0}{\left(1 - e_0^2\right)^{3/2}}\left[-\left(\frac{\bar{r}}{a_0}\cos \bar{f} + 2e_0\right) + \sqrt{1 - e_0^2} \cdot \frac{\nu}{1+\nu}\frac{d}{dl}\frac{\bar{r}}{a_0}\sin \bar{f} + \left(\frac{h^2}{h_0^2} - 1\right) e_0 \frac{\bar{r}}{a_0}\frac{\sin \bar{f}}{\sqrt{1 - e_0^2}}\frac{d}{dl}\frac{\bar{r}}{a_0}\cos \bar{f}\right] . \quad (23)$$

Equation 14 takes a very simple form:

$$\frac{d}{dt}\frac{h_0}{h} = M_3 \frac{\partial \Omega}{\partial l} + N_3 \cdot r \frac{\partial \Omega}{\partial r}, \tag{24}$$

$$M_3 = \frac{n_0 a_0}{1 - e_0^2} \frac{\bar{r}^2}{a_0^2}, \tag{25}$$

$$N_3 = - \frac{n_0 a_0}{1 - e_0^2} \cdot \frac{\bar{r}}{a_0} \cdot \frac{e_0 \sin \bar{f}}{\sqrt{1 - e_0^2}}. \tag{26}$$

Equations 18, 21, and 24 bear more resemblance to the equations from Hansen's planetary theory than to the corresponding equations from his lunar theory, and they can be used to serve as a basis of a purely numerical development. The value of $n_0 y$ must be determined in such a way that no purely secular term is present in equation 21.

PERTURBATIONS OF THE POSITION OF THE ORBIT PLANE

In previous work[4] the author has established the following equations for the determination of λ parameters:

$$\frac{d\lambda_1}{dt} = + n_0 \alpha \lambda_2 + \frac{1}{2} hr \frac{\partial \Omega}{\partial Z} \left[+ \lambda_4 \cos (\bar{f} + \omega) - \lambda_3 \sin (\bar{f} + \omega) \right], \tag{27}$$

$$\frac{d\lambda_2}{dt} = - n_0 \alpha \lambda_1 + \frac{1}{2} hr \frac{\partial \Omega}{\partial Z} \left[- \lambda_3 \cos (\bar{f} + \omega) - \lambda_4 \sin (\bar{f} + \omega) \right], \tag{28}$$

$$\frac{d\lambda_3}{dt} = + n_0 \eta \lambda_4 + \frac{1}{2} hr \frac{\partial \Omega}{\partial Z} \left[+ \lambda_2 \cos (\bar{f} + \omega) + \lambda_1 \sin (\bar{f} + \omega) \right], \tag{29}$$

$$\frac{d\lambda_4}{dt} = - n_0 \eta \lambda_3 + \frac{1}{2} hr \frac{\partial \Omega}{\partial Z} \left[- \lambda_1 \cos (\bar{f} + \omega) + \lambda_2 \sin (\bar{f} + \omega) \right]. \tag{30}$$

We have

$$\frac{\partial \Omega}{\partial Z} = - m' \left(\frac{1}{\Delta^3} - \frac{1}{r'^3} \right) r' \sin (v' - \psi') \sin I. \tag{31}$$

Taking the relations

$$v' = \bar{f}' + \pi_0' + n_0 y' t$$

and

$$\psi' = \psi_0' - n_0 (\alpha + \eta) t + K - N$$

into account, we obtain from equation 31:

$$\frac{1}{2}\frac{\partial\Omega}{\partial Z} = -m'\left(\frac{1}{\Delta^3}-\frac{1}{r'^3}\right)r'\left[\left(\lambda_2\lambda_4-\lambda_1\lambda_3\right)\cos\left(\bar{f}'+\omega'\right)+\left(\lambda_1\lambda_4+\lambda_2\lambda_3\right)\sin\left(\bar{f}'+\omega'\right)\right]. \quad (32)$$

We have from equation 7:

$$\frac{1}{2}\frac{\partial S}{\partial\lambda_1} = +\lambda_1\cos\left(\bar{f}+\bar{f}'+\omega+\omega'\right)-\lambda_2\sin\left(\bar{f}+\bar{f}'+\omega+\omega'\right), \quad (33)$$

$$\frac{1}{2}\frac{\partial S}{\partial\lambda_2} = -\lambda_2\cos\left(\bar{f}+\bar{f}'+\omega+\omega'\right)-\lambda_1\sin\left(\bar{f}+\bar{f}'+\omega+\omega'\right), \quad (34)$$

$$\frac{1}{2}\frac{\partial S}{\partial\lambda_3} = -\lambda_3\cos\left(\bar{f}-\bar{f}'+\omega-\omega'\right)-\lambda_4\sin\left(\bar{f}-\bar{f}'+\omega-\omega'\right), \quad (35)$$

$$\frac{1}{2}\frac{\partial S}{\partial\lambda_4} = +\lambda_4\cos\left(\bar{f}-\bar{f}'+\omega-\omega'\right)-\lambda_3\sin\left(\bar{f}-\bar{f}'+\omega-\omega'\right). \quad (36)$$

Substituting equation 32 into equations 27 to 30 and taking equations 33 to 36 into account, we deduce the system of equations that replace Hansen's equations for his P, Q, and K parameters:

$$\frac{d\lambda_1}{dt} = +n_0\alpha\lambda_2+\frac{1}{4}\frac{h}{h_0}\cdot\frac{a_0 n_0}{\sqrt{1-e_0^2}}\left[+\left(\lambda_4^2+\lambda_3^2\right)\frac{\partial\Omega}{\partial\lambda_2}-\left(\lambda_1\lambda_4+\lambda_2\lambda_3\right)\frac{\partial\Omega}{\partial\lambda_3}-\left(\lambda_2\lambda_4-\lambda_1\lambda_3\right)\frac{\partial\Omega}{\partial\lambda_4}\right], \quad (37)$$

$$\frac{d\lambda_2}{dt} = -n_0\alpha\lambda_1+\frac{1}{4}\frac{h}{h_0}\cdot\frac{a_0 n_0}{\sqrt{1-e_0^2}}\left[-\left(\lambda_4^2+\lambda_3^2\right)\frac{\partial\Omega}{\partial\lambda_1}-\left(\lambda_2\lambda_4-\lambda_1\lambda_3\right)\frac{\partial\Omega}{\partial\lambda_3}+\left(\lambda_1\lambda_4+\lambda_2\lambda_3\right)\frac{\partial\Omega}{\partial\lambda_4}\right], \quad (38)$$

$$\frac{d\lambda_3}{dt} = +n_0\eta\lambda_4+\frac{1}{4}\frac{h}{h_0}\cdot\frac{a_0 n_0}{\sqrt{1-e_0^2}}\left[-\left(\lambda_1^2+\lambda_2^2\right)\frac{\partial\Omega}{\partial\lambda_4}+\left(\lambda_1\lambda_4+\lambda_2\lambda_3\right)\frac{\partial\Omega}{\partial\lambda_1}+\left(\lambda_2\lambda_4-\lambda_1\lambda_3\right)\frac{\partial\Omega}{\partial\lambda_2}\right], \quad (39)$$

$$\frac{d\lambda_4}{dt} = -n_0\eta\lambda_3+\frac{1}{4}\frac{h}{h_0}\cdot\frac{a_0 n_0}{\sqrt{1-e_0^2}}\left[+\left(\lambda_1^2+\lambda_2^2\right)\frac{\partial\Omega}{\partial\lambda_3}+\left(\lambda_2\lambda_4-\lambda_1\lambda_3\right)\frac{\partial\Omega}{\partial\lambda_1}-\left(\lambda_1\lambda_4+\lambda_2\lambda_3\right)\frac{\partial\Omega}{\partial\lambda_2}\right]. \quad (40)$$

If necessary, the effect of the secular motion of the disturbing body's orbit can be easily taken into account. We have for this effect[6]:

$$\frac{dI}{dt} = -\frac{di'}{dt}\cos\left(\psi'-\sigma'\right)-\frac{d\Omega'}{dt}\sin i'\sin\left(\psi'-\sigma'\right), \quad (41)$$

$$\frac{dN}{dt} = +\frac{1}{2}\left[-\frac{di'}{dt}\sin\left(\psi'-\sigma'\right)+\frac{d\Omega'}{dt}\sin i'\cos\left(\psi'-\sigma'\right)\right]\operatorname{ctg}\frac{1}{2}I, \quad (42)$$

$$\frac{dK}{dt} = +\frac{1}{2}\left[-\frac{di'}{dt}\sin(\psi'-\sigma') + \frac{d\Omega'}{dt}\sin i'\cos(\psi'-\sigma')\right]\operatorname{tg}\frac{1}{2}I \cdot \tag{43}$$

We can neglect the effect of change on i' on the motion of the satellite. Substituting

$$\psi' = \psi_0' - n_0(\alpha+\eta) + K - N$$

into equations 41 to 43 and putting

$$\theta = (\pi_0' + n_0 y' t - \sigma') - \omega' ,$$

we obtain the effect of the secular motion of the disturbing body's orbit plane on the relative position of the satellite's orbit plane in the form:

$$\frac{d\delta\lambda_1}{dt} = -\frac{1}{2}(+\lambda_3\cos\theta + \lambda_4\sin\theta)\cdot\sin i'\frac{d\Omega'}{dt} , \tag{44}$$

$$\frac{d\delta\lambda_2}{dt} = +\frac{1}{2}(+\lambda_4\cos\theta - \lambda_3\sin\theta)\cdot\sin i'\frac{d\Omega'}{dt} , \tag{45}$$

$$\frac{d\delta\lambda_3}{dt} = +\frac{1}{2}(+\lambda_1\cos\theta + \lambda_2\sin\theta)\cdot\sin i'\frac{d\Omega'}{dt} , \tag{46}$$

$$\frac{d\delta\lambda_4}{dt} = -\frac{1}{2}(+\lambda_2\cos\theta - \lambda_1\sin\theta)\cdot\sin i'\frac{d\Omega'}{dt} . \tag{47}$$

The combined effect of direct perturbations with the secular motion of the orbital plane of the disturbing body is given by the system of equations representing the sum of equations 37 to 40 and of 44 to 47.

$$\frac{d\lambda_1}{dt} = +n_0\alpha\lambda_2 + \frac{1}{4}\frac{h}{h_0}\cdot\frac{a_0 n_0}{\sqrt{1-e_0^2}}\left[+(+\lambda_4^2+\lambda_3^2)\frac{\partial\Omega}{\partial\lambda_2} - (+\lambda_1\lambda_4+\lambda_2\lambda_3)\frac{\partial\Omega}{\partial\lambda_3} - (+\lambda_2\lambda_4-\lambda_1\lambda_3)\frac{\partial\Omega}{\partial\lambda_4}\right]$$
$$-\frac{1}{2}(+\lambda_3\cos\theta+\lambda_4\sin\theta)\cdot\sin i'\frac{d\Omega'}{dt} , \tag{48}$$

$$\frac{d\lambda_2}{dt} = -n_0\alpha\lambda_1 + \frac{1}{4}\frac{h}{h_0}\cdot\frac{a_0 n_0}{\sqrt{1-e_0^2}}\left[-(\lambda_4^2+\lambda_3^2)\frac{\partial\Omega}{\partial\lambda_1} - (+\lambda_2\lambda_4-\lambda_1\lambda_3)\frac{\partial\Omega}{\partial\lambda_3} + (+\lambda_1\lambda_4+\lambda_2\lambda_3)\frac{\partial\Omega}{\partial\lambda_4}\right]$$
$$+\frac{1}{2}(+\lambda_4\cos\theta-\lambda_3\sin\theta)\cdot\sin i'\frac{d\Omega'}{dt} , \tag{49}$$

$$\frac{d\lambda_3}{dt} = +n_0\eta\lambda_4 + \frac{1}{4}\frac{h}{h_0}\cdot\frac{a_0 n_0}{\sqrt{1-e_0^2}}\left[-(\lambda_1^2+\lambda_2^2)\frac{\partial\Omega}{\partial\lambda_4} + (+\lambda_1\lambda_4+\lambda_2\lambda_3)\frac{\partial\Omega}{\partial\lambda_1} + (+\lambda_2\lambda_4-\lambda_1\lambda_3)\frac{\partial\Omega}{\partial\lambda_2}\right]$$
$$+\frac{1}{2}(+\lambda_1\cos\theta+\lambda_2\sin\theta)\cdot\sin i'\frac{d\Omega'}{dt} , \tag{50}$$

$$\frac{d\lambda_4}{dt} = -n_0\eta\lambda_3 + \frac{1}{4}\frac{h}{h_0}\cdot\frac{a_0 n_0}{\sqrt{1-e_0^2}}\left[+(\lambda_1^2+\lambda_2^2)\frac{\partial\Omega}{\partial\lambda_3} + (+\lambda_2\lambda_4-\lambda_1\lambda_3)\frac{\partial\Omega}{\partial\lambda_1} - (+\lambda_1\lambda_4+\lambda_2\lambda_3)\frac{\partial\Omega}{\partial\lambda_2}\right]$$

$$-\frac{1}{2}(+\lambda_2\cos\theta-\lambda_1\sin\theta)\cdot\sin i'\frac{d\Omega'}{dt}\cdot \qquad (51)$$

The values of α and η must be determined in such a way that no terms linear in time are present in equations 49 and 50.

DETERMINATION OF CONSTANTS OF INTEGRATION

The series for Ψ has the form

$$\Psi = \sum S\sin(ig+i'g'+j\omega+j'\omega'), \qquad (52)$$

and no additive constant of integration appears in this development. The secular motion $n_0 y$ of the perigee of the satellite in its orbit plane is determined in such a way that no constant term is present in equation 21 for $d\Psi/dt$ and, consequently, no secular term appears in equation 52. The series for

$$\frac{h_0}{h} \text{ and } \Upsilon$$

both will have the form

$$\sum C\cos(ig+i'g'+j\omega+j'\omega'), \qquad (53)$$

and an additive constant of integration is associated with the series of the form 53. Let

$$\left[\frac{h_0}{h}\right], [\Upsilon], \text{ and } [\Psi]$$

be the series obtained by the formal integration procedure as applied to equations 18, 21, and 24. We then have

$$\frac{h_0}{h} = 1 + c_1 + \left[\frac{h_0}{h}\right] = 1 + \Delta, \qquad (54)$$

$$\Upsilon = c_2 + [\Upsilon], \qquad (55)$$

$$\Psi = [\Psi]. \qquad (56)$$

At this point it is convenient to introduce Hansen's Ξ and W functions. We have[1]

$$\Xi = -1 - \frac{h_0}{h} + 2\frac{h}{h_0} - \frac{3}{2}e_0\Upsilon. \qquad (57)$$

Taking equations 54 and 55 into account, we can write the last equation in the form

$$\Xi = -3\Delta - \frac{3}{2}e_0\Upsilon + 2(\Delta^2-\Delta^3+\cdots) \qquad (58)$$

or in the form

$$\Xi = \left(-3c_1 - \frac{3}{2}e_0 c_2\right) + [\Xi], \qquad (59)$$

where we put

$$[\Xi] = -3\left[\frac{h_0}{h}\right] - \frac{3}{2}e_0[\Upsilon] + 2(\Delta^2-\Delta^3+\cdots). \qquad (60)$$

The value for Δ in the last equation can be taken from the previous approximation.

We put with Hansen:

$$\bar{W}_0 = \Xi + \Upsilon\left(\frac{\bar{\rho}}{a_0}\cos\bar{\varphi} + \frac{3}{2}e_0\right) + \Psi\frac{\bar{\rho}}{a_0}\sin\bar{\varphi} \qquad (61)$$

$$\bar{W} = \Xi + \Upsilon\left(\frac{\bar{r}}{a_0}\cos\bar{f} + \frac{3}{2}e_0\right) + \Psi\frac{\bar{r}}{a_0}\sin\bar{f}. \qquad (62)$$

We have:

$$\bar{W} - \bar{W}_0 = T\left(\frac{\bar{r}}{a_0}\cos\bar{f} - \frac{\bar{\rho}}{a_0}\cos\bar{\varphi}\right) + \Psi\left(\frac{\bar{r}}{a_0}\sin\bar{f} - \frac{\bar{\rho}}{a_0}\sin\bar{\varphi}\right), \tag{63}$$

$$\left(\frac{\bar{r}}{a_0}\cos\bar{f}\right) - \left(\frac{\bar{\rho}}{a_0}\cos\bar{\varphi}\right) = \sum_n \frac{(n_0\,\delta z)^n}{n!}\frac{d^n}{dg^n}\frac{\bar{\rho}}{a_0}\cos\bar{\varphi}, \tag{64}$$

$$\left(\frac{\bar{r}}{a_0}\sin\bar{f}\right) - \left(\frac{\bar{\rho}}{a_0}\sin\bar{\varphi}\right) = \sum_n \frac{(n_0\,\delta z)^n}{n!}\frac{d^n}{dg^n}\frac{\bar{\rho}}{a_0}\sin\bar{\varphi}; \tag{65}$$

in addition,

$$\left(\frac{r}{a_0}\right)^2 - \left(\frac{\bar{\rho}}{a_0}\right)^2 = \sum_n \frac{(n_0\,\delta z)^n}{n!}\frac{d^n}{dg^n}\frac{\bar{\rho}^2}{a_0^2}. \tag{66}$$

The value of $n_0\,\delta z$ taken from the previous approximation can be used in the computation of the right-hand sides of equations 63 to 66 for our purposes.

The generalized Hill's formula[7]

$$\frac{dn_0\,\delta z}{dt} = n_0\frac{\bar{W} + \nu^2}{1 - \nu^2} - \frac{n_0\,y}{\sqrt{1 - e_0^2}}\frac{\bar{r}^2}{a_0^2} \tag{67}$$

can be written in the form

$$\frac{dn_0\,\delta z}{dt} = n_0\bar{W} + \frac{n_0\,\nu^2(\bar{W} + 1)}{1 - \nu^2} - \frac{n_0\,y}{\sqrt{1 - e_0^2}}\frac{\bar{r}^2}{a_0^2} \tag{68}$$

or in the form

$$\frac{dn_0\,\delta z}{dt} = n_0\bar{W}_0 - \frac{n_0\,y}{\sqrt{1 - e_0^2}}\frac{\bar{\rho}^2}{a_0^2} + B, \tag{69}$$

where

$$B = n_0\,(\bar{W} - \bar{W}_0) - \frac{n_0\,y}{\sqrt{1 - e_0^2}}\left(\frac{\bar{r}^2}{a_0^2} - \frac{\bar{\rho}^2}{a_0^2}\right) + \frac{n_0\,\nu^2(\bar{W} + 1)}{1 - \nu^2}. \tag{70}$$

The value of B is computed on the basis of the previous approximation. Designating by $[\bar{W}_0]$ the value of $\bar{W}_0$ as obtained using the formal integration procedure, we have

$$\bar{W}_0 = +\left(-3c_1 - \frac{3}{2}e_0\,c_2\right) + c_2\left(\frac{\bar{\rho}}{a_0}\cos\bar{\varphi} + \frac{3}{2}e_0\right) + [\bar{W}_0], \tag{71}$$

$$[\bar{W}_0] = [\Xi] + [\Upsilon]\left(\frac{\bar{\rho}}{a_0}\cos\bar{\varphi} + \frac{3}{2}e_0\right) + [\Psi]\frac{\bar{\rho}}{a_0}\sin\bar{\varphi}\ ; \tag{72}$$

and equation 69 takes the form

$$\frac{dn_0\,\delta z}{dt} = n_0\left(-3c_1 - \frac{3}{2}e_0 c_2\right) + n_0 c_2\left(\frac{\bar{\rho}}{a_0}\cos\bar{\varphi} + \frac{3}{2}e_0\right) - \frac{n_0 y}{\sqrt{1-e_0^2}}\frac{\bar{\rho}^2}{a_0^2} + n_0[\bar{W}_0] + B\ . \tag{73}$$

Two constants of integration, c_1 and c_2, must be adjusted in such a way that no secular term and no term with the argument g is present in $n_0\delta z$. The part

$$n_0[\bar{W}_0] + B - \frac{n_0 y}{\sqrt{1-e_0^2}}\frac{\bar{\rho}^2}{a_0^2}$$

contains the term of the form

$$A_1 + A_2\cos g\ .$$

Let β be the coefficient of the term in $(\bar{\rho}/a_0)\cos\bar{\varphi}+\frac{3}{2}e_0$ having the argument g. No constant term is present in $(\bar{\rho}/a_0)\cos\bar{\varphi}+\frac{3}{2}e_0$. Thus we must have

$$n_0\left(-3c_1 - \frac{3}{2}e_0 c_2\right) + A_1 = 0\ ,$$

$$n_0 c_2\beta + A_2 = 0\ ;$$

and

$$c_2 = -\frac{A_2}{\beta n_0}\ ,$$

$$c_1 = +\frac{1}{3}\left(\frac{A_1}{n_0} + \frac{3}{2}\frac{e_0 A_2}{\beta n_0}\right).$$

After the constants c_1 and c_2 are determined, we obtain the perturbations of the mean anomaly by integrating equation 69 or 73. The perturbations of the radius vector are obtained from the formula

$$\bar{W} = -1 + 2\frac{h_0}{h}\cdot\frac{1}{1+\nu} - \frac{h_0}{h}\ , \tag{74}$$

which leads to the formula

$$\nu = \frac{1}{2}(\Delta - \bar{W}) - \frac{1}{2}(\Delta + \bar{W})\nu\ , \tag{75}$$

which is convenient for the use of the iteration process. This formula was used by the author in the artificial satellites theory.[4]

Determination of the integration constants associated with the λ parameters is done in exactly the same way as in reference 4. We put

$$\lambda_1 = \sin\frac{1}{2}I_0 + \frac{1}{2}(A+B) + [\lambda_1]\ , \tag{76}$$

$$\lambda_2 = [\lambda_2]\ , \tag{77}$$

$$\lambda_3 = [\lambda_3]\ , \tag{78}$$

$$\lambda_4 = \cos\frac{1}{2}I_0 + \frac{1}{2}(A-B) + [\lambda_4]\ , \tag{79}$$

where $[\lambda_i]$ $(i=1,2,3,4)$ are the series obtained by the formal integration procedure as applied to equations 37 to 40 and where A and B are the constants of integration. We omit here the development leading to the equations for the determination of A and B, because this development can be found in reference 4, and give only the final results:

$$A^2 + 2A\left(\cos\frac{1}{2}I_0 + \sin\frac{1}{2}I_0\right) + (11) = 0\ , \tag{80}$$

$$B^2 - 2B\left(\cos\frac{1}{2}I_0 - \sin\frac{1}{2}I_0\right) + (12) = 0\ , \tag{81}$$

where

$$(11) = \text{const. part in } \left\{\left([\lambda_1] + [\lambda_4]\right)^2 + \left([\lambda_2] - [\lambda_3]\right)^2\right\}$$

and

$$(12) = \text{const. part in } \left\{\left([\lambda_1] - [\lambda_4]\right)^2 + \left([\lambda_2] + [\lambda_3]\right)^2\right\} .$$

DETERMINATION OF THE POSITION AND VELOCITY VECTORS

We determine the position of the satellite with respect to the moving system of coordinates, whose x axis is in the orbit plane of the disturbing body and is directed toward the mean position of the pericenter; the y axis is in the orbit plane of the disturbing body, normal to the x axis; and the z axis is normal, in the positive direction, to the x and y axes. Introducing the matrices

$$A_1(\alpha) = \begin{bmatrix} +1 & 0 & 0 \\ 0 & +\cos\alpha & -\sin\alpha \\ 0 & +\sin\alpha & +\cos\alpha \end{bmatrix},$$

$$A_3(\alpha) = \begin{bmatrix} +\cos\alpha & -\sin\alpha & 0 \\ +\sin\alpha & +\cos\alpha & 0 \\ 0 & 0 & +1 \end{bmatrix},$$

we have

$$\mathbf{r} = A_3\left(\psi' - \pi_0' - n_0 y' t\right) \cdot A_1(I) \cdot A_3\left(\pi_0 + n_0 y t - \psi\right)(1+\nu) \begin{bmatrix} \bar{r}\cos\bar{f} \\ \bar{r}\sin\bar{f} \\ 0 \end{bmatrix} .$$

Substituting

$$\psi = \psi_0 + n_0(\eta - \alpha) - K - N ,$$

$$\psi' = \psi_0' - n_0(\eta + \alpha) + K - N$$

into the last equation, we deduce

$$\mathbf{r} = A_3(-\omega') \cdot \Lambda \cdot A_3(\omega)(1+\nu) \begin{bmatrix} \bar{r}\cos\bar{f} \\ \bar{r}\sin\bar{f} \\ 0 \end{bmatrix} , \tag{82}$$

where

$$\bar{r}\cos\bar{f} = a_0\left(\cos E - e_0\right) ,$$

$$\bar{r}\sin\bar{f} = a_0\sqrt{1-e_0^2}\sin E ,$$

$$E - e_0\sin E = g_0 + n_0 t + n_0\,\delta z ;$$

where (Reference 4)

$$\lambda_{11} = +\lambda_1^2 - \lambda_2^2 - \lambda_3^2 + \lambda_4^2 ,$$

$$\lambda_{12} = -2\left(\lambda_3\lambda_4 + \lambda_1\lambda_2\right) ,$$

$$\lambda_{13} = +2\left(\lambda_1\lambda_3 - \lambda_2\lambda_4\right) ,$$

$$\lambda_{21} = +2\left(\lambda_3\lambda_4 - \lambda_1\lambda_2\right) ,$$

$$\lambda_{22} = -\lambda_1^2 + \lambda_2^2 - \lambda_3^2 + \lambda_4^2 ,$$

$$\lambda_{23} = -2\left(\lambda_1\lambda_4 + \lambda_2\lambda_3\right) ,$$

$$\lambda_{31} = +2\left(\lambda_3\lambda_1 + \lambda_2\lambda_4\right) ,$$

$$\lambda_{32} = +2\left(\lambda_1\lambda_4 - \lambda_2\lambda_3\right) ,$$

$$\lambda_{33} = -\lambda_1^2 - \lambda_2^2 + \lambda_3^2 + \lambda_4^2 .$$

For the absolute velocity, but decomposed along the axes of the moving system (x,y,z), we have

$$\mathbf{v} = h\,A_3\left(\psi' - \pi_0' - n_0\,y' t\right)\cdot A_1\,(I)\cdot A_3\,(\chi-\psi)\cdot\begin{bmatrix} -\sin f \\ \cos f + e \\ 0 \end{bmatrix} .$$

The last equation can be written in the form

$$\mathbf{v} = h A_3(-\omega') \cdot \Lambda \cdot A_3(\omega) \cdot A_3(\chi - \pi_0 - n_0 yt) \cdot \begin{bmatrix} -\sin f \\ \cos f + e \\ 0 \end{bmatrix} . \tag{83}$$

Taking the equation

$$f + \chi - \pi_0 - n_0 yt = \bar{f}$$

into account, equation 83 can be reduced to the form

$$\mathbf{v} = h A_3(-\omega') \cdot \Lambda \cdot A_3(\omega) \cdot \begin{bmatrix} -\sin \bar{f} - e \sin(\chi - \pi_0 - n_0 yt) \\ +\cos \bar{f} + e \cos(\chi - \pi_0 - n_0 yt) \\ 0 \end{bmatrix} . \tag{84}$$

Taking equations 12 and 13 into account, we have finally:

$$\mathbf{v} = \frac{a_0 n_0}{\sqrt{1 - e_0^2}} A_3(-\omega') \cdot \Lambda \cdot A_3(\omega) \cdot \begin{bmatrix} -\frac{h}{h_0} \sin \bar{f} & -\frac{1 - e_0^2}{2} \Psi \\ +\frac{h}{h_0}(\cos \bar{f} + e_0) + \frac{1 - e_0^2}{2} \Upsilon & \\ 0 & \end{bmatrix} . \tag{85}$$

Equation 85 is analogous to equation 22 of the article published by A. Bailie and R. Bryant[8] on the determination of the velocity of artificial satellites. The transformation from a moving system (x,y,z) to the inertial system does not represent any difficulty and therefore is omitted.

COLLECTION OF FORMULAS

Input information:

$$g_0,\ \omega_0,\ \omega_0',\ I_0,\ a_0,\ e_0,\ n_0 ,$$

$$g_0',\ a',\ e',\ n',\ i',\ \Omega' .$$

Basic arguments:

$$g = g_0 + n_0 t , \qquad g' = g_0' + n_0' t ,$$

$$\omega = \omega_0 + n_0 (y + \alpha - \eta) t ,$$

$$\omega' = \omega_0' + n_0 (\alpha + \eta + y') t ,$$

$$\theta = (\pi_0' + n_0 y' t - \sigma') - \omega' .$$

We make use of the following standard formulas of the elliptic motion:

$$\frac{\rho}{a_0}\cos\varphi = -\frac{3}{2}e_0 + 2\sum_{p=1}^{+\infty}\frac{J_p'(pe_0)}{p}\cos p\gamma ,$$

$$\frac{\rho}{a_0}\sin\varphi = \frac{2\sqrt{1-e_0^2}}{e_0}\sum_{p=1}^{+\infty}\frac{J_p(pe_0)}{p}\sin p\gamma ,$$

$$\frac{\rho}{a_0} = 1 + \frac{1}{2}e_0^2 - 2e_0\sum_{p=1}^{+\infty}\frac{J_p'(pe_0)}{p}\cos p\gamma ,$$

$$\frac{\rho^2}{a_0^2} = 1 + \frac{3}{2}e_0^2 - 4\sum_{p=1}^{+\infty}\frac{J_p(pe_0)}{p^2}\cos p\gamma ,$$

$$\frac{a_0}{\rho} = 1 + 2\sum_{p=1}^{+\infty}J_p(pe_0)\cos p\gamma ,$$

$$\frac{a'}{r'} = 1 + 2\sum_{p=1}^{+\infty}J_p(pe')\cos pg'$$

If preferred, Cayley's tables can be used instead of formulas given above.

$$s_1 = (1+\nu)\cdot\frac{\rho}{a_0}\cdot\frac{a'}{r'}\cos(\varphi+\bar{f}'+\omega+\omega') ,$$

$$s_2 = (1+\nu)\cdot\frac{\rho}{a_0}\cdot\frac{a'}{r'}\sin(\varphi+\bar{f}'+\omega+\omega') ,$$

$$s_3 = (1+\nu)\cdot\frac{\rho}{a_0}\cdot\frac{a'}{r'}\cdot\cos(\varphi-\bar{f}'+\omega-\omega') ,$$

$$s_4 = (1+\nu)\cdot\frac{\rho}{a_0}\cdot\frac{a'}{r'}\cdot\sin(\varphi-\bar{f}'+\omega-\omega') ,$$

$$s = +(\lambda_1^2-\lambda_2^2)s_1 - 2\lambda_1\lambda_2 s_2 + (\lambda_4^2-\lambda_3^2)s_3 - 2\lambda_3\lambda_4 s_4 ;$$

$$p = (1+\nu)\frac{\rho}{a_0}\cdot\frac{a'}{r'} ;$$

$$\frac{1}{2}\frac{\partial s}{\partial\lambda_1} = +\lambda_1 s_1 - \lambda_2 s_2 = \sigma_1 ,$$

$$\frac{1}{2}\frac{\partial s}{\partial\lambda_2} = -\lambda_2 s_1 - \lambda_1 s_2 = \sigma_2 ,$$

$$\frac{1}{2}\frac{\partial s}{\partial\lambda_3} = -\lambda_3 s_3 - \lambda_4 s_4 = \sigma_3 ,$$

$$\frac{1}{2}\frac{\partial s}{\partial\lambda_4} = +\lambda_4 s_3 - \lambda_3 s_4 = \sigma_4 ;$$

$$M_1 = \frac{2a_0 n_0}{1 - e_0^2}\left[\frac{1}{e_0}\left(1 - e_0^2 - \frac{\rho^2}{a_0^2}\right) - \frac{\nu}{1+\nu}\frac{1}{e_0}\left(1 - e_0^2 - \frac{\rho}{a_0}\right) + \left(\frac{h^2}{h_0^2} - 1\right)\frac{1}{e_0}\frac{\rho}{a_0}\left(1 - \frac{\rho}{a_0}\right)\right]$$

$$N_1 = \frac{2a_0 n_0}{1 - e_0^2} \cdot \frac{\rho}{a_0} \cdot \frac{\sin\varphi}{\sqrt{1-e_0^2}}\left[1 - \frac{a_0}{\rho} \cdot \frac{\nu}{1+\nu} - \left(\frac{h^2}{h_0^2} - 1\right)\left(\frac{a_0}{\rho} - 1\right)\right],$$

$$M_2 = \frac{2a_0 n_0}{1 - e_0^2}\left[\frac{1}{\sqrt{1-e_0^2}}\int\left(2\frac{\rho}{a_0}\cos\varphi + 3e_0\right)d\gamma - \frac{\nu}{1+\nu}\frac{\rho}{a_0}\sin\varphi + \left(\frac{h^2}{h_0^2} - 1\right)\frac{\rho^2}{a_0^2} \cdot \frac{\sin\varphi}{1-e_0^2}\right],$$

$$N_2 = \frac{2a_0 n_0}{(1-e_0^2)^{3/2}}\left[-\left(\frac{\rho}{a_0}\cos\varphi + 2e_0\right) + \sqrt{1-e_0^2}\,\frac{\nu}{1+\nu}\frac{d}{d\gamma}\frac{\rho}{a_0}\sin\varphi + \left(\frac{h^2}{h_0^2} - 1\right)e_0\frac{\rho}{a_0}\frac{\sin\varphi}{\sqrt{1-e_0^2}}\frac{d}{d\gamma}\frac{\rho}{a_0}\cos\varphi\right],$$

$$M_3 = +\frac{n_0 a_0}{1 - e_0^2}\frac{\rho^2}{a_0^2},$$

$$N_3 = -\frac{n_0 a_0}{1 - e_0^2} \cdot \frac{\rho}{a_0} \cdot \frac{e_0 \sin\varphi}{\sqrt{1-e_0^2}};$$

$$\Omega_1 = \frac{m' a^2}{a'^3} \cdot \frac{a'}{r'} \cdot \left(\frac{3}{2}s^2 - \frac{1}{2}p^2\right),$$

$$\Omega_2 = \frac{m' a^3}{a'^4} \cdot \frac{a'}{r'} \cdot \left(\frac{5}{2}s^3 - \frac{3}{2}sp^2\right),$$

$$\Omega_3 = \frac{m' a^4}{a'^5} \cdot \frac{a'}{r'} \cdot \left(\frac{35}{8}s^4 - \frac{15}{4}s^2p^2 + \frac{3}{8}p^4\right),$$

. . . . . .

$$\Omega = \Omega_1 + \Omega_2 + \Omega_3 + \cdots,$$

$$\frac{\partial\Omega}{\partial\gamma} = \frac{\partial\Omega_1}{\partial\gamma} + \frac{\partial\Omega_2}{\partial\gamma} + \frac{\partial\Omega_3}{\partial\gamma} + \cdots,$$

$$\rho \frac{\partial \Omega}{\partial \rho} = 2\Omega_1 + 3\Omega_2 + 4\Omega_3 + \cdots ,$$

$$\frac{a}{r} \cdot \frac{r'}{a'^2} \cdot \frac{\partial \Omega}{\partial S} = +\frac{m' a^2}{a'^3} \cdot \frac{a'}{r'^2} \cdot 3s + \frac{m' a^3}{a'^4} \cdot \frac{a'}{r'^2} \cdot \left(\frac{15}{2} s^2 - \frac{3}{2} \rho^2\right) + \frac{m' a^4}{a'^5} \cdot \frac{a'}{r'^2} \cdot \left(\frac{35}{2} s^3 - \frac{15}{2} s \rho^2\right) + \cdots ,$$

$$T_i = M_i \frac{\partial \Omega}{\partial \gamma} + N_i \cdot \rho \frac{\partial \Omega}{\partial \rho} \qquad (i = 1,2,3) ;$$

$$F_i = \sum_n \frac{1}{n!} \frac{\overline{\partial^n T_i}}{\partial \gamma^n} (n_0 \delta z)^n \qquad (n = 0,1,2,3, \cdots) ,$$

$$\frac{d\Upsilon}{dt} = +n_0 y \Psi + F_1 ,$$

$$\frac{d\Psi}{dt} = -n_0 y \left(\Upsilon + 2 \frac{h}{h_0} \cdot \frac{e_0}{1 - e_0^2}\right) + F_2 ,$$

$$\frac{d}{dt} \frac{h_0}{h} = F_3 ;$$

$$G_1 = \frac{1}{2} \frac{h}{h_0} \cdot \frac{a_0 n_0}{\sqrt{1 - e_0^2}} \cdot \left(\frac{a}{\rho} \cdot \frac{\bar{r}'}{a'^2} \cdot \frac{1}{1 + \nu} \frac{\partial \Omega}{\partial S}\right) \cdot \left[+ (\lambda_3^2 + \lambda_4^2) \sigma_2 - (\lambda_1 \lambda_4 + \lambda_2 \lambda_3) \sigma_3 - (\lambda_2 \lambda_4 - \lambda_1 \lambda_3) \sigma_4\right] ,$$

$$G_2 = \frac{1}{2} \frac{h}{h_0} \cdot \frac{a_0 n_0}{\sqrt{1 - e_0^2}} \cdot \left(\frac{a}{\rho} \cdot \frac{\bar{r}'}{a'^2} \cdot \frac{1}{1 + \nu} \frac{\partial \Omega}{\partial S}\right) \left[- (\lambda_4^2 + \lambda_3^2) \sigma_1 - (\lambda_2 \lambda_4 - \lambda_1 \lambda_3) \sigma_3 + (\lambda_1 \lambda_4 + \lambda_2 \lambda_3) \sigma_4\right] ,$$

$$G_3 = \frac{1}{2} \frac{h}{h_0} \cdot \frac{a_0 n_0}{\sqrt{1 - e_0^2}} \cdot \left(\frac{a}{\rho} \cdot \frac{\bar{r}'}{a'^2} \cdot \frac{1}{1 + \nu} \frac{\partial \Omega}{\partial S}\right) \left[- (\lambda_1^2 + \lambda_2^2) \sigma_4 + (\lambda_1 \lambda_4 + \lambda_2 \lambda_3) \sigma_1 + (\lambda_2 \lambda_4 - \lambda_1 \lambda_3) \sigma_2\right] ,$$

$$G_4 = \frac{1}{2}\frac{h}{h_0}\cdot\frac{a_0 n_0}{\sqrt{1-e_0^2}}\cdot\left(\frac{a}{\rho}\cdot\frac{\bar{r}'}{a'}\cdot\frac{1}{1+\nu}\frac{\partial\Omega}{\partial S}\right)\cdot\Big[+\left(\lambda_1^2+\lambda_2^2\right)\sigma_3 + \left(\lambda_2\lambda_4-\lambda_1\lambda_3\right)\sigma_1 - \left(\lambda_1\lambda_4+\lambda_2\lambda_3\right)\sigma_2\Big];$$

$$H_i = \sum_n \frac{1}{n!}\left(n_0\,\delta z\right)^n \frac{\overline{\partial^n G_i}}{\partial\gamma^n} \qquad (i = 0,1,2,3,4),$$

$$\frac{d\lambda_1}{dt} = +n_0\,\alpha\lambda_2 + H_1 - \frac{1}{2}\left(+\lambda_4\sin\theta+\lambda_3\cos\theta\right)\cdot\sin i'\,\frac{d\Omega'}{dt},$$

$$\frac{d\lambda_2}{dt} = -n_0\,\alpha\lambda_1 + H_2 + \frac{1}{2}\left(+\lambda_4\cos\theta-\lambda_3\sin\theta\right)\cdot\sin i'\,\frac{d\Omega'}{dt},$$

$$\frac{d\lambda_3}{dt} = +n_0\,\eta\lambda_4 + H_3 + \frac{1}{2}\left(+\lambda_1\cos\theta+\lambda_2\sin\theta\right)\cdot\sin i'\,\frac{d\Omega'}{dt},$$

$$\frac{d\lambda_4}{dt} = -n_0\,\eta\lambda_3 + H_4 + \frac{1}{2}\left(+\lambda_1\sin\theta-\lambda_2\cos\theta\right)\cdot\sin i'\,\frac{d\Omega'}{dt}.$$

Let $\left[\frac{h_0}{h}\right]$, $[\Upsilon]$, $[\Psi]$, $[\lambda_i]$ $(i = 1,2,3,4)$

be the series obtained by the formal integration process.

$$\frac{h_0}{h} = 1 + c_1 + \left[\frac{h_0}{h}\right] = 1 + \Delta,$$

$$\Upsilon = c_2 + [\Upsilon],$$

$$\Psi = [\Psi],$$

$$[\Xi] = -3\left[\frac{h_0}{h}\right] - \frac{3}{2}e_0\,[\Upsilon] + 2\left(\Delta^2-\Delta^3+\cdots\right),$$

$$\Xi = -3\Delta - \frac{3}{2}e_0\Upsilon + 2\left(\Delta^2-\Delta^3+\cdots\right);$$

$$\left(\frac{\bar{r}}{a_0}\cos\bar{f}\right) - \left(\frac{\bar{\rho}}{a_0}\cos\bar{\varphi}\right) = \sum_n \frac{\left(n_0\,\delta z\right)^n}{n!}\frac{d^n}{dg^n}\frac{\bar{\rho}}{a_0}\cos\bar{\varphi},$$

$$\left(\frac{\bar{r}}{a_0}\sin\bar{f}\right)-\left(\frac{\bar{\rho}}{a_0}\sin\bar{\varphi}\right)=\sum_n\frac{(n_0\,\delta z)^n}{n!}\frac{d^n}{dg^n}\frac{\bar{\rho}}{a_0}\sin\bar{\varphi}\,,$$

$$\left(\frac{\bar{r}}{a_0}\right)^2-\left(\frac{\bar{\rho}}{a_0}\right)^2=\sum_n\frac{(n_0\,\delta z)^n}{n!}\frac{d^n}{dg^n}\frac{\bar{\rho}^2}{a_0^2}\;;$$

$$\left[\bar{W}_0\right]=[\Xi]+[\Upsilon]\left(\frac{\bar{\rho}}{a_0}\cos\bar{\varphi}+\frac{3}{2}e_0\right)+[\Psi]\frac{\bar{\rho}}{a_0}\sin\bar{\varphi}\;;$$

$$B=n_0\,[\Upsilon]\left(\frac{\bar{r}}{a_0}\cos\bar{f}-\frac{\bar{\rho}}{a_0}\cos\bar{\varphi}\right)+[\Psi]\left(\frac{\bar{r}}{a_0}\sin\bar{f}-\frac{\bar{\rho}}{a_0}\sin\bar{\varphi}\right)$$

$$-\frac{n_0\,y}{\sqrt{1-e_0^2}}\left(\frac{\bar{r}^2}{a_0^2}-\frac{\bar{\rho}^2}{a_0^2}\right)+\frac{n_0\,\nu^2(1+\bar{W})}{1-\nu^2}\,,$$

$$\frac{dn_0\,\delta z}{dt}=n_0\left(-3c_1-\frac{3}{2}e_0\,c_2\right)+n_0\,c_2\left(\frac{\bar{\rho}}{a_0}\cos\bar{\varphi}+\frac{3}{2}e_0\right)$$

$$-\frac{n_0\,y}{\sqrt{1-e_0^2}}\frac{\bar{\rho}^2}{a_0^2}+n_0\left[\bar{W}_0\right]+B\,,$$

$$n_0\left[\bar{W}_0\right]-\frac{n_0\,y}{\sqrt{1-e_0^2}}\cdot\frac{\bar{\rho}^2}{a_0^2}+3=A_1+A_2\cos g+\cdots\,,$$

$$\frac{\bar{\rho}}{a_0}\cos\bar{\varphi}+\frac{3}{2}e_0=\beta\cos g+\cdots\;;$$

$$c_1=\frac{A_1}{3n_0}+\frac{A_2\,e_0}{2\beta n_0}\,,$$

$$c_2=-\frac{A_2}{\beta n_0}\;;$$

$$\bar{W}=\Xi+\Upsilon\left(\frac{\bar{r}}{a_0}\cos\bar{f}+\frac{3}{2}e_0\right)+\Psi\,\frac{\bar{r}}{a_0}\sin\bar{f}\,,$$

$$\nu=\frac{1}{2}(\Lambda-\bar{W})-\frac{1}{2}(\Lambda+\bar{W})\,\nu\;;$$

$$(11) = \text{const. in } \left\{\left(\left[\lambda_1\right] + \left[\lambda_4\right]\right)^2 + \left(\left[\lambda_2\right] - \left[\lambda_3\right]\right)^2\right\},$$

$$(12) = \text{const. in } \left\{\left(\left[\lambda_1\right] - \left[\lambda_4\right]\right)^2 + \left(\left[\lambda_2\right] + \left[\lambda_3\right]\right)^2\right\},$$

$$A^2 + 2A\left(\cos\frac{1}{2}I_0 + \sin\frac{1}{2}I_0\right) + (11) = 0,$$

$$B^2 - 2B\left(\cos\frac{1}{2}I_0 - \sin\frac{1}{2}I_0\right) + (12) = 0;$$

$$\lambda_1 = \sin\frac{1}{2}I_0 + \frac{1}{2}(A+B) + \left[\lambda_1\right],$$

$$\lambda_2 = \left[\lambda_2\right],$$

$$\lambda_3 = \left[\lambda_3\right],$$

$$\lambda_4 = \cos\frac{1}{2}I_0 + \frac{1}{2}(A-B) + \left[\lambda_4\right];$$

$$A_1(\alpha) = \begin{bmatrix} +1 & 0 & 0 \\ 0 & +\cos\alpha & -\sin\alpha \\ 0 & +\sin\alpha & +\cos\alpha \end{bmatrix},$$

$$A_3(\alpha) = \begin{bmatrix} +\cos\alpha & -\sin\alpha & 0 \\ +\sin\alpha & +\cos\alpha & 0 \\ 0 & 0 & +1 \end{bmatrix},$$

$$\mathbf{r} = A_3(-\omega') \cdot \Lambda \cdot A_3(\omega)(1+\nu) \begin{bmatrix} \bar{r}\cos\bar{f} \\ \bar{r}\sin\bar{f} \\ 0 \end{bmatrix},$$

$$\mathbf{v} = \frac{a_0 n_0}{\sqrt{1-e_0^2}} A_3(-\omega') \cdot \Lambda \cdot A_3(\omega) \cdot \begin{bmatrix} -\frac{h}{h_0}\sin\bar{f} & -\frac{1}{2}(1-e_0^2)\,\Psi \\ +\frac{h}{h_0}\left(\cos\bar{f}+e_0\right) & +\frac{1}{2}(1-e_0^2)\,\Upsilon \\ 0 & \end{bmatrix};$$

$$\lambda_{11} = +\lambda_1^2 + \lambda_2^2 - \lambda_3^2 + \lambda_4^2,$$

$$\lambda_{12} = -2(\lambda_3\lambda_4 + \lambda_1\lambda_2),$$

$$\lambda_{13} = +2(\lambda_1\lambda_3 - \lambda_2\lambda_4),$$

$$\lambda_{21} = +2(\lambda_3\lambda_4 - \lambda_1\lambda_2),$$

$$\lambda_{22} = -\lambda_1^2 + \lambda_2^2 - \lambda_3^2 + \lambda_4^2,$$

$$\lambda_{23} = -2(\lambda_1\lambda_4 + \lambda_2\lambda_3),$$

$$\lambda_{31} = +2(\lambda_1\lambda_3 + \lambda_2\lambda_4),$$

$$\lambda_{32} = +2(\lambda_1\lambda_4 - \lambda_2\lambda_3),$$

$$\lambda_{33} = -\lambda_1^2 - \lambda_2^2 + \lambda_3^2 + \lambda_4^2.$$

CONCLUSION

A systematic investigation of the orbital stability of satellites under different initial conditions can be based on the development given herein. In light of recent results concerning the stability of lunar satellites by G. Chebotarev and S. Kirpichnikov,[9] which were obtained by the method of numerical integration, it would be of interest to investigate the problem using the development into periodic series. We can expect that one of the most influential long period terms in the coordinates will be the term with the period equal to the sidereal revolution of the earth around the moon. The described method will be programed by the Theoretical Division, Goddard Space Flight Center, for computation on the IBM 7090. We have to expect that the number of terms in the development either for planetary satellites (with the exception of the moon) or for lunar satellites will not be large, and thus the computation can be done with the limited accuracy.

REFERENCES

1. HANSEN, P. A., "Fundamenta Nova Investigationis Orbitae Verae quam Luna Perlustrat," Gotha: C. Glaeser, 1838.
2. LIMEKHOVA, E. N., "Application of Delaunay's Method to the Motion of Lunar Artificial Satellites," *Byulleten Instituta Teoreticheskoy Astronomii*, **8**(7): 512–529, 1962 (in Russian).
3. HANSEN, P. A., "Auseinandersetzung Einer Zweckmassigen Metode zur Berechnung der Absoluten Störungen der Kleinen Planeten. Erste Abhandlung; Zweite Abhandlung," *Abhandlungen der Mathematisch-Physischen Classe der Königlich Sächsischen Gesellschaft der Wissenschaften*, **3**, 41–218, 1857; **4**, 1–148, 1859.
4. MUSEN, P., "Application of Hansen's Theory to the Motion of an Artificial Satellite in the Gravitational Field of the Earth," *J. Geophys. Res.*, **64**(12): 2271–2279, December 1959.
5. CAYLEY, A., "Tables of the Developments of Functions in the Theory of Elliptic Motion," *Mem. Roy. Astronom. Soc.*, **29**, 191–306, 1859.
6. HANSEN, P. A., "Darlegung der Theoretischen Berechnung der in den Mondtafeln Angewandten Störungen. Erste Abhandlung," *Abhandlungen der Mathematisch-Physischen Classe der Königlich Sächsischen Gesellschaft der Wissenschaften*, **6**, 91–497, 1862.
7. HILL, G. W., "Note on Hansen's General Formulae for Perturbations," *Amer. J. Math.*, **4**, 256–259, 1881.
8. BAILIE, A., and BRYANT, R., "Osculating Elements Derived from the Modified Hansen Theory for the Motion of an Artificial Satellite," *Astronom. J.*, **65**(8): 451–453, October 1960.
9. CHEBOTAREV, G. A., and KIRPICHNIKOV, S. N., "On the Problem of Stability of Motion of an Artificial Lunar Satellite," *Byulleten Instituta Teoreticheskoy Astronomii*, **8**(6): 402–404, 1962 (in Russian)

Appendix A

NOTATIONS

a_0	the mean semi-major axis of the satellite's orbit
a, e, n	the osculating elements of the satellite
e_0	the mean eccentricity of the satellite's orbit
f	the osculating true anomaly
$\bar{f}'$	the true anomaly of the disturbing body
$g=g_0+n_0t$	the undisturbed mean anomaly of the satellite
g_0	satellite's mean anomaly at the epoch
h	$=an/(1-e^2)^{1/2}$
h_0	$=a_0n_0/(1-e_0^2)^{1/2}$
I_0	the mean value of the mutual inclination of two orbit planes
$-2K$	the periodic part in $\psi-\psi'$
$l=g_0+n_0t+n_0\delta z$	the perturbed mean anomaly of the satellite
m'	the mass of the disturbing body
$-2N$	the periodic part in $\psi+\psi'$
n_0	the mean anomalistic mean motion of the satellite (a_0 is defined as $n_0^{-2/3}$; the gravitational constant is put equal to 1)
$n_0\delta z$	the perturbations in the satellite's mean anomaly
$-2n_0\alpha t$	the secular part in $\psi+\psi'$
$+2n_0\eta t$	the secular part in $\psi-\psi'$
$\mathbf{r}$	the position vector of the satellite, $\mathbf{r}=\|\mathbf{r}\|$
$\mathbf{r}'$	the position vector of the disturbing body, $\|\mathbf{r}'\|=\mathbf{r}'=\bar{r}'$
$\bar{r}, \bar{f}$	the radius vector and the true anomaly depending upon l, $\bar{r}\cos\bar{f}=a_0(\cos E-e_0)$ $\bar{r}\sin\bar{f}=a_0(1-e_0^2)^{1/2}\sin E$ $E-e_0\sin E=l$
S	$=\cos(\mathbf{r}, \mathbf{r}')$
v	the true orbital longitude of the satellite, reckoned from the departure point in the satellite's orbit plane
v'	the true orbital longitude of the disturbing body, reckoned from the departure point in the body's orbit plane
γ	the auxiliary mean anomaly; γ is replaced by l at the later stage of computation
Δ	the distance between the satellite and the disturbing body
$\lambda_1, \lambda_2, \lambda_3, \lambda_4$	the parameters defining the periodic perturbations of the orbit plane, $\lambda_1=\sin\frac{1}{2}\cos N$, $\lambda_2=\sin\frac{1}{2}\sin N$, $\lambda_3=\cos\frac{1}{2}\sin K$, $\lambda_4=\cos\frac{1}{2}\cos K$
ν	the perturbations in the radius vector $\mathbf{r}$, $\mathbf{r}=(1+\nu)\delta$
π_0+n_0yt	the mean orbital longitude of the pericenter of the satellite
$\pi'_0+n_0y't$	the mean orbital longitude of the pericenter of the disturbing body

ρ, φ	the radius vector and the true anomaly depending upon the auxiliary mean anomaly γ, $\rho \cos \varphi = a_0 (\cos \epsilon - e_0)$ $\rho \sin \varphi = a_0(1-e_0^2)^{1/2} \sin \epsilon$ $\epsilon - e_0 \sin \epsilon = \gamma$
$\bar{\rho}, \bar{\varphi}$	the radius vector and the true anomaly depending upon g, $\bar{\rho} \cos \bar{\varphi} = a_0 (\cos \bar{\epsilon} - e_0)$ $\bar{\rho} \sin \bar{\varphi} = a_0(1-e_0^2)^{1\,2} \sin \bar{\epsilon}$ $\bar{\epsilon} - e_0 \sin \bar{\epsilon} = g$
σ'	the true orbital longitude of the node of the disturbing body's orbit plane; the fixed ecliptic and equinox serve as a reference
χ	the true orbital longitude of the osculating pericenter of the satellite
ψ	the true orbital longitude of the common node, as reckoned from the departure point in the orbit plane of the satellite
ψ'	the true orbital longitude of the common node, as reckoned from the departure point in the orbit plane of the disturbing body
ψ_0, ψ_0'	the constant parts in ψ and ψ'
Ω	the disturbing function
$\partial\Omega/\partial r$	the radial component of the disturbing force
$\partial\Omega/\partial Z$	the component of the disturbing force normal to the orbit plane

ON THE MOTION OF A 24-HOUR SATELLITE

PETER MUSEN AND ANN E. BAILIE

Goddard Space Flight Center

The theory and conditions for stability of a satellite with a 24-hour period are given. Bohlin's resonance theory was applied to obtain the solution. It is shown that the integrals of the problem can be represented in series form, with respect to the small parameter w, which is proportional to the mean motion of the critical argument in a nonresonance case. Expressions for the period of libration and the mean motion of the critical argument in the unstable case are also given. A system of formulas is presented which can be used to compute any particular case.

INTRODUCTION

In this paper the authors have investigated the stability conditions of a satellite with a period of revolution approximately equal to 1 day. The criteria for stability are obtained in a form valid for large inclinations—provided that the 8 power of the eccentricity is negligible.

THE DISTURBING FUNCTION

The disturbing function consists of a secular part produced by the zonal harmonics k_2 and k_4, and a periodic part produced by the ellipticity of the earth's equator. The periodic disturbing function has the form

$$F_1 = \frac{3}{2}\frac{\mu A_{22}}{r^5}\left(x^2 - y^2\right) . \tag{1}$$

The x-axis is directed along the semi-major of the equator, and the z-axis is directed along the axis of rotation of the earth. Substituting

$$\frac{x}{r} = \frac{1+\cos i}{2}\cos\left(f+\omega+\Omega-n't\right) + \frac{1-\cos i}{2}\cos\left(f+\omega-\Omega+n't\right) ,$$

$$\frac{y}{r} = \frac{1+\cos i}{2}\sin\left(f+\omega+\Omega-n't\right) - \frac{1-\cos i}{2}\sin\left(f+\omega-\Omega+n't\right) ,$$

$$\frac{z}{r} = \sin\left(f+\omega\right)\sin i$$

into equation 1, we obtain

$$F_1 = \frac{3}{8}\frac{\mu A_{22}}{a^3}\left(1+\cos i\right)^2\left(\frac{a}{r}\right)^3\cos\left(2f+2\omega+2\Omega-2n't\right)$$

$$+\frac{3}{4}\frac{\mu A_{22}}{a^3}\sin^2 i\left(\frac{a}{r}\right)^3\cos\left(2\Omega-2n't\right)$$

$$+\frac{3}{8}\frac{\mu A_{22}}{a^3}\left(1-\cos i\right)^2\left(\frac{a}{r}\right)^3\cos\left(2f+2\omega-2\Omega+2n't\right) . \tag{2}$$

Developing equation 2 into a series in terms of the mean anomaly l with coefficients developed in powers of the eccentricity, and retaining the long period terms only, we deduce

$$F_1 = Q_0 \cos(2l + 2\omega + 2\Omega - 2n't) + Q_1 \cos(2l + 2\Omega - 2n't) + Q_2 \cos(2l - 2\omega + 2\Omega - 2n't) , \tag{3}$$

where

$$Q_0 = \frac{3}{8}\frac{\mu A_{22}}{a^3}(1+\cos i)^2\left(1 - \frac{5}{2}e^2 + \frac{13}{16}e^4 - \frac{35}{288}e^6\right) ,$$

$$Q_1 = \frac{3}{4}\frac{\mu A_{22}}{a^3}\sin^2 i\left(\frac{9}{4}e^2 + \frac{7}{4}e^4 + \frac{141}{64}e^6\right) ,$$

$$Q_2 = \frac{3}{8}\frac{\mu A_{22}}{a^3}(1-\cos i)^2\left(\frac{1}{24}e^4 + \frac{7}{240}e^6\right) .$$

If the mean motion of the satellite is such that it causes the satellite to remain above a particular longitude of the earth for some time, the first term in equation 3 is the most significant, and will be treated in accordance with the theory of resonance. The last two terms will produce only small, long-period terms in the osculating elements.

The coefficient A_{22} is approximately of the same order as the coefficient of the fourth zonal harmonic, and consequently, the secular part F_0 of the disturbing function need not be developed beyond the results established by Brouwer.[1] Adding the term $n'[\mu a(1-e^2)]^{1/2}\cos i$, produced by the rotation of the earth to Brouwer's development, we have

$$F_0 = \frac{\mu^2}{2L^2} + n'H + \frac{\mu^4 k_2}{L^3 G^3}\left(-\frac{1}{2} + \frac{3}{2}\frac{H^2}{G^2}\right)$$

$$+ \frac{\mu^6 k_4}{L^{10}}\left(\frac{15}{16}\frac{L^7}{G^7} - \frac{9}{16}\frac{L^5}{G^5}\right)\left(1 - 10\frac{H^2}{G^2} + \frac{35}{3}\frac{H^4}{G^4}\right)$$

$$+ \frac{\mu^6 k_2^2}{L^{10}}\left[+\frac{15}{32}\frac{L^5}{G^5}\left(1 - \frac{18}{5}\frac{H^2}{G^2} + \frac{H^4}{G^4}\right)\right.$$

$$\left. + \frac{3}{8}\frac{L^6}{G^6}\left(1 - 6\frac{H^2}{G^2} + 9\frac{H^4}{G^4}\right) - \frac{15}{32}\frac{L^7}{G^7}\left(1 - 2\frac{H^2}{G^2} - 7\frac{H^4}{G^4}\right)\right] ,$$

where L, G, and H are Delaunay variables

$$L = \sqrt{\mu a} , \quad G = \sqrt{\mu a(1-e^2)} , \quad H = \sqrt{\mu a\,1-e^2)}\cos i ,$$

and k_2 and k_4 are the coefficients of the second and fourth zonal harmonics, respectively.

Considering the form of the main argument, it is more convenient to use the canonical set of Poincare rather than that of Delaunay.

Thus

$$\left.\begin{aligned} x_1 &= \sqrt{\mu a} = L, & y_1 &= l + \omega + \Omega - n't, \\ x_2 &= \sqrt{\mu a}\left(1 - \sqrt{1-e^2}\right) = L - G, \text{ and} & y_2 &= -\omega, \text{ and} \\ x_3 &= \sqrt{\mu a}\left(1 - \sqrt{1-e^2}\cos i\right) = L - H, & y_3 &= -\Omega + n't. \end{aligned}\right\} \tag{4}$$

The complete Hamiltonian is

$$F = F_0 + F_1, \tag{5}$$

but the disturbing function can be expressed more conveniently by means of the auxiliary quantities x_1, ϵ, and γ, where $\epsilon = x_2/x_1$, $\gamma = 1 - x_3/x_1$. And the following relations exist:

$$L = x_1,$$
$$G/L = 1 - \epsilon,$$
$$H/G = \cos i = \gamma/(1-\epsilon),$$

and

$$e^2 = 2\epsilon - \epsilon^2.$$

The following expressions were used in the actual computations:

$$\begin{aligned} F_0 = {} & n'\gamma x_1 + \frac{\mu^2}{2x_1^2} \\ & + \frac{\mu^4 k_2}{2x_1^6}\left[-\left(1 + 3\epsilon + 6\epsilon + 10\epsilon^3\right) + \gamma^2\left(3 + 15\epsilon + 45\epsilon^2 + 105\epsilon^3\right)\right] \\ & + \frac{\mu^6 k_2^2}{x_1^{10}}\left[\frac{3}{32}\left(4 + 14\epsilon + 19\epsilon^2 - 21\epsilon^3\right) - \frac{3}{16}\gamma^2\left(16 + 114\epsilon + 459\epsilon^2 + 1371\epsilon^3\right)\right. \\ & \qquad \left. + \frac{3}{32}\gamma^4\left(76 + 790\epsilon + 4515\epsilon^2 + 18755\epsilon^3\right)\right] \\ & + \frac{\mu^6 k_4}{x_1^{10}}\left[\frac{3}{16}\left(2 + 20\epsilon + 95\epsilon^2 + 315\epsilon^3\right) - \frac{15}{8}\gamma^2\left(2 + 24\epsilon + 141\epsilon^2 + 573\epsilon^3\right)\right. \\ & \qquad \left. + \frac{5}{16}\gamma^4\left(14 + 196\epsilon + 1365\epsilon^2 + 6545\epsilon^3\right)\right], \end{aligned}$$

$$Q_0 = \frac{\mu^4 A_{22}}{x_1^6}\left[\frac{1}{96}\left(36 - 180\epsilon + 207\epsilon^2 - 152\epsilon^3\right) + \frac{1}{48}\gamma\left(36 - 144\epsilon + 63\epsilon^2 - 89\epsilon^3\right)\right.$$

$$\left. + \frac{1}{96}\gamma^2\left(36 - 108\epsilon - 45\epsilon^2 - 134\epsilon^3\right)\right],$$

$$Q_1 = \frac{\mu^4 A_{22}}{x_1^6}\left[\frac{3}{32}\left(36\epsilon + 38\epsilon^2 + 85\epsilon^3\right) - \frac{3}{32}\gamma^2\left(36\epsilon + 110\epsilon^2 + 269\epsilon^3\right)\right],$$

$$Q_2 = \frac{\mu^4 A_{22}}{x_1^6}\left[\frac{1}{80}\left(5\epsilon^2 + 2\epsilon^3\right) - \frac{1}{40}\gamma\left(5\epsilon^2 + 7\epsilon^3\right) + \frac{1}{80}\gamma^2\left(5\epsilon^2 + 12\epsilon^3\right)\right].$$

The first and second derivatives of F_0 and Q_0 with respect to x_1 can be obtained by using the expression

$$F' = \frac{\partial F}{\partial x_1} + \frac{1-\gamma}{x_1}\frac{\partial F}{\partial \gamma} - \frac{\epsilon}{x_1}\frac{\partial F}{\partial \epsilon},$$

from which we obtain:

$$F_0' = n' - \frac{\mu^2}{x_1^3} + \frac{3\mu^4 k_2}{2x_1^7}\left[\left(2 + 7\epsilon + 16\epsilon^2 + 30\epsilon^3\right) + 2\gamma\left(1 + 5\epsilon + 15\epsilon^2 + 35\epsilon^3\right) - \gamma^2\left(8 + 45\epsilon + 150\epsilon^2 + 385\epsilon^3\right)\right]$$

$$- \frac{\mu^6 k_2^2}{x_1^{11}}\left[\frac{3}{32}\left(40 + 154\epsilon + 228\epsilon^2 - 273\epsilon^3\right) + \frac{3}{8}\gamma\left(16 + 114\epsilon + 459\epsilon^2 + 1371\epsilon^3\right)\right.$$

$$- \frac{9}{16}\gamma^2\left(64 + 494\epsilon + 2142\epsilon^2 + 6855\epsilon^3\right) - \frac{3}{8}\gamma^3\left(76 + 790\epsilon + 4515\epsilon^2 + 18755\epsilon^3\right)$$

$$\left. + \frac{3}{32}\gamma^4\left(1064 + 11850\epsilon + 72240\epsilon^2 + 318835\epsilon^3\right)\right]$$

$$- \frac{\mu^6 k_4}{x_1^{11}}\left[\frac{15}{16}\left(4 + 44\epsilon + 228\epsilon^2 + 819\epsilon^3\right) + \frac{15}{4}\gamma\left(2 + 24\epsilon + 141\epsilon^2 + 573\epsilon^3\right)\right.$$

$$- \frac{45}{8}\gamma^2\left(8 + 104\epsilon + 658\epsilon^2 + 2865\epsilon^3\right) - \frac{35}{4}\gamma^3\left(2 + 28\epsilon + 195\epsilon^2 + 935\epsilon^3\right)$$

$$\left. + \frac{35}{16}\gamma^4\left(28 + 420\epsilon + 3120\epsilon^2 + 15895\epsilon^3\right)\right];$$

$$F_0'' = + \frac{3\mu^2}{x_1^4} - \frac{\mu^4 k_2}{x_1^8}\left[3\left(6+23\epsilon+57\epsilon^2+115\epsilon^3\right) + 6\gamma\left(8+45\epsilon+150\epsilon^2+385\epsilon^3\right) - 9\gamma^2\left(12+75\epsilon+275\epsilon^2+770\epsilon^3\right)\right]$$

$$+ \frac{\mu^6 k_2^2}{x_1^{12}}\left[\frac{3}{16}\left(188+696\epsilon+564\epsilon^2-4653\epsilon^3\right) + \frac{9}{8}\gamma\left(128+988\epsilon+4284\epsilon^2+13710\epsilon^3\right)\right.$$

$$- \frac{9}{8}\gamma^2\left(340+2668\epsilon+11550\epsilon^2+36085\epsilon^3\right) - \frac{3}{4}\gamma^3\left(1064+11850\epsilon+72240\epsilon^2+318835\epsilon^3\right)$$

$$\left. + \frac{45}{16}\gamma^4\left(532+6320\epsilon+40936\epsilon^2+191301\epsilon^3\right)\right]$$

$$+ \frac{\mu^6 k_4}{x_1^{12}}\left[\frac{45}{8}\left(6+72\epsilon+400\epsilon^2+1529\epsilon^3\right) + \frac{45}{2}\gamma\left(8+104\epsilon+658\epsilon^2+2865\epsilon^3\right)\right.$$

$$- \frac{15}{4}\gamma^2\left(142+1988\epsilon+13440\epsilon^2+62215\epsilon^3\right) - \frac{35}{2}\gamma^3\left(28+420\epsilon+3120\epsilon^2+15895\epsilon^3\right)$$

$$\left. + \frac{525}{8}\gamma^4\left(14+224\epsilon+1768\epsilon^2+9537\epsilon^3\right)\right];$$

$$Q_0' = - \frac{\mu^4 A_{22}}{x_1^7}\left[\frac{1}{48}\left(72-486\epsilon+765\epsilon^2-595\epsilon^3\right) + \frac{3}{4}\gamma\left(6-29\epsilon+17\epsilon^2-21\epsilon^3\right)\right.$$

$$\left. + \frac{1}{48}\gamma^2\left(144-486\epsilon-225\epsilon^2-737\epsilon^3\right)\right];$$

$$Q_0'' = + \frac{\mu^4 A_{22}}{x_1^8}\left[\frac{1}{48}\left(288-2844\epsilon+6273\epsilon^2-5194\epsilon^3\right) + \frac{1}{24}\gamma\left(720-4212\epsilon+3285\epsilon^2-3421\epsilon^3\right)\right.$$

$$\left. + \frac{1}{16}\gamma^2\left(432-1620\epsilon-825\epsilon^2-2948\epsilon^3\right)\right].$$

If only the secular and the first, most important, periodic terms of the disturbing function are retained, we have:

$$F = F_0 + Q_0 \cos 2y_1 . \tag{6}$$

The libration points and the points lying on the intersection of two branches of the separatrix are determined from the equations

$$\frac{\partial F}{\partial x_1} = \frac{\partial F_0}{\partial x_1} + \frac{\partial Q_0}{\partial x_1}\cos 2y_1$$

and

$$\frac{\partial F}{\partial y_1} = - 2\, Q_0 \sin 2y_1$$

if the values of x_2 and x_3 are fixed. From equation 6 we deduce that

$$\frac{\partial F_0}{\partial x_1} + \frac{\partial Q_0}{\partial x_1} = 0 \tag{7}$$

$$\text{for } y_1 = 0, \pi, \tag{8}$$

$$\frac{\partial F_0}{\partial x_1} - \frac{\partial Q_0}{\partial x_1} = 0 \tag{9}$$

$$\text{for } y_1 = \frac{\pi}{2}, \frac{3\pi}{2}. \tag{10}$$

The values $y = \pi/2, 3\pi/2$ reduce the disturbing function (equation 5) to a minimum. Consequently, they correspond to the stability position and determine the libration points. The set $y_1 = 0, \pi$ gives the points on the separatrix and corresponds to unstable positions. For satellites moving in the equatorial plane the stable positions are on the equator's minor axis, and the unstable positions are on the major axis.[2] It is convenient in the complete problem, as defined by equation 5, to retain the development around the characteristic points defined by equations 7 through 10, and to write the disturbing function in the form

$$F = R_0 + R_1 ,$$

where

$$R_0 = F_0 - Q_0 ,$$

$$R_1 = 2Q_0 \cos^2 y_1 + Q_1 \cos\left(2y_1 + 2y_2\right) + Q_2 \cos\left(2y_1 + 4y_2\right)$$

for the stable case, and

$$R_0 = F_0 + Q_0 , \tag{11}$$

$$R_1 = -2Q_0 \sin^2 y_1 + Q_1 \cos\left(2y_1 + 2y_2\right) + Q_2 \cos\left(2y_1 + 4y_2\right) \tag{12}$$

for the unstable case.

The canonical transformation (equation 4) removes the time and the argument y_3 from the disturbing function. Consequently, this problem contains the energy integral

$$R_0 + R_1 = -C , \tag{13}$$

and the integral

$$x_3 = \sqrt{\mu a}\left(1 - \sqrt{1 - e^2}\cos i\right) = \alpha_3$$

is constant.

THE STABLE CASE

By substituting $x_1=\partial S/\partial y_1$, and $x_2=\partial S/\partial y_2$ into equation 13, we transform it into a Hamilton-Jacobi partial differential equation. Letting

$$S = S_0 + S_1 (y_1) + S_2 (y_1 , y_2) + S_3 (y_1 , y_2) \cdots ,$$

$$C = C_0 + C_1 + C_2 + C_3 + \cdots ,$$

and

$$S_0 = \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3 ,$$

where α_1 and α_2 can be considered as two constants of integration. We replace the partial differential equation 13 by the system of Bohlin's equations.[3] These may be deduced by developing the left-hand side of equation 13 into a Taylor's series about α_1, α_2, α_3:

$$R_0 (\alpha_1, \alpha_2, \alpha_3) = - C_0 , \tag{14}$$

$$w_{01} \frac{\partial S_2}{\partial y_2} + \frac{1}{2} w_{20} \left(\frac{\partial S_1}{\partial y_1}\right)^2 + w_{10} \frac{\partial S_1}{\partial y_1} + 2Q_0 \cos^2 y_1 + Q_1 \cos (2y_1 + 2y_2) + Q_2 \cos (2y_1 + 4y_2) = 0 , \tag{15}$$

$$w_{01} \frac{\partial S_3}{\partial y_2} + \left(w_{10} + w_{20} \frac{\partial S_1}{\partial y_1}\right) \frac{\partial S_2}{\partial y_1} + \frac{\partial S_1}{\partial y_1} \frac{\partial S_2}{\partial y_2} w_{11} + \frac{1}{6} w_{30} \left(\frac{\partial S_1}{\partial y_1}\right)^3$$

$$+ \left[2Q_0' \cos^2 y_1 + Q_1' \cos (2y_1 + 2y_2) + Q_2' \cos (2y_1 + 4y_2)\right] \frac{\partial S_1}{\partial y_1} = - C_1 , \tag{16}$$

where

$$w_{ij} = \frac{\partial^{i+j} R_0 (\alpha_1, \alpha_2, \alpha_3)}{\partial \alpha_1^i \quad \partial \alpha_2^j}$$

and

$$Q_i' = \frac{\partial Q_i (\alpha_1, \alpha_2, \alpha_3)}{\partial \alpha_1}$$

Equation 14 can be considered as a defining equation for C_0.

Imposing on our solution an additional condition—that no secular term with respect to y_2 is contained in S_1, S_2, S_3, . . . , but only in S_0—we deduce from equations 14, 15, and 16 that

$$S_1 = \phi_1 (y_1) ,$$

$$S_2 = \phi_2 (y_1) - \frac{Q_1}{2w_{01}} \sin (2y_1 + 2y_2) - \frac{Q_2}{4w_{01}} \sin (2y_1 + 4y_2) ,$$

and also

$$\frac{1}{2} w_{20} \phi_1'^2 + w_{10} \phi_1' + 2Q_0 \cos^2 y_1 = 0 , \tag{17}$$

$$\phi_2' (w_{10} + w_{20} \phi_1') + \frac{1}{6} w_{30} \phi_1'^3 + 2\phi_1' Q_0' \cos^2 y_1 = - C_1 . \tag{18}$$

From equation 17, we have the standard first approximation

$$\phi_1' = -w + A, \tag{19}$$

where

$$w = w_{10}/w_{20},$$

and

$$A = \pm \sqrt{w^2 - \frac{4Q_0}{w_{20}} \cos^2 y_1} . \tag{20}$$

If

$$\frac{4Q_0}{w_{20} w^2} > 1 , \tag{21}$$

then cos y_1 oscillates between the limits $-(w/2)(w_{20}/Q_0),^{1/2}+(w/2)(w_{20}/Q_0).^{1/2}$ This condition (equation 21) can be written in the form:

$$(F_0' - Q_0')^2 - 4Q_0(F_0'' - Q_0'') < 0, \tag{22}$$

(where, in this case, the primes represent differentiation with respect to α_1), together with the condition

$$\cos^2 y_1 < \frac{(F_0' - Q_0')^2}{4Q_0 (F_0'' - Q_0'')}$$

which, for a certain moment of time, must be fulfilled for the motion to be stable. By eliminating ϕ_1' and $\cos^2 y_1$ from equation 18 by means of equations 19 and 20, we can write an expression for ϕ_2':

$$\phi_2' = - \left(\frac{1}{6} \frac{w_{30}}{w_{20}} - \frac{1}{2} \frac{Q_0'}{Q_0}\right) A^2 + \frac{1}{2} \left(\frac{w_{30}}{w_{20}} - \frac{Q_0'}{Q_0}\right) Aw - \frac{1}{2} \left(\frac{w_{30}}{w_{20}} + \frac{Q_0'}{Q_0}\right) w^2 - \frac{1}{A} \left[C_1 - \left(\frac{1}{6} \frac{w_{30}}{w_{20}} + \frac{1}{2} \frac{Q_0'}{Q_0}\right) w^3\right] .$$

The divisor A in the last term may become zero. Therefore, to remove a source of possible discontinuity, we set

$$C_1 = \left(\frac{1}{6} \frac{w_{30}}{w_{20}} + \frac{1}{2} \frac{Q_0'}{Q_0}\right) w^3 , \tag{23}$$

and ϕ_2' becomes a polynomial in A which, taking equation 20 into account, can also be written as

$$\phi_2' = - \frac{2}{3} w^2 \frac{w_{30}}{w_{20}} + \left(\frac{1}{3} \frac{w_{30}}{w_{20}} - \frac{Q_0'}{Q_0}\right) \frac{Q_0}{w_{20}} + \left(\frac{1}{3} \frac{w_{30}}{w_{20}} - \frac{Q_0'}{Q_0}\right) \frac{Q_0}{w_{20}} \cos 2y_1 + \frac{1}{2} \left(\frac{w_{30}}{w_{20}} - \frac{Q_0'}{Q_0}\right) wA .$$

Setting

$$A^{(1)} = -w - \frac{2}{3} w \frac{w_{30}}{w_{20}} + \left(\frac{1}{3} \frac{w_{30}}{w_{20}} - \frac{Q_0'}{Q_0} \right) \frac{Q_0}{w_{20}} , \tag{24}$$

$$A^{(2)} = \left[w + \frac{1}{2} \left(\frac{w_{30}}{w_{20}} - \frac{Q_0'}{Q_0} \right) w^2 \right]^2 , \tag{25}$$

$$A^{(3)} = \frac{4Q_0}{w_{20}} \left[1 + \frac{1}{2} \left(\frac{w_{30}}{w_{20}} - \frac{Q_0'}{Q_0} \right) w \right]^2 > 0 , \tag{26}$$

$$A^{(4)} = \frac{Q_0}{w_{20}} \left(\frac{1}{6} \frac{w_{30}}{w_{20}} - \frac{1}{2} \frac{Q_0'}{Q_0} \right), \tag{27}$$

$$A^{(5)} = -\frac{Q_1}{2w_{01}} , \tag{28}$$

$$A^{(6)} = -\frac{Q_2}{4w_{01}} , \tag{29}$$

$$\frac{\partial A^{(i)}}{\partial \alpha_j} = A_j^{(i)} ,$$

$$\frac{\partial C}{\partial \alpha_j} = n_j ,$$

$$\frac{1}{2} \left(A_j^{(2)} - A^{(2)} \frac{A_j^{(3)}}{A^{(3)}} \right) = M_j , \tag{30}$$

we obtain the Hamiltonian function S with its integrals in the form:

$$S = \left(\alpha_1 + A^{(1)} \right) y_1 + \alpha_2 y_2 + \alpha_3 y_3 \pm \int \sqrt{A^{(2)} - A^{(3)} \cos^2 y_1} \, dy_1$$
$$+ A^{(4)} \sin 2y_1 + A^{(5)} \sin \left(2y_1 + 2y_2 \right) + A^{(6)} \sin \left(2y_1 + 4y_2 \right) ;$$

and

$$x_1 = \frac{\partial S}{\partial y_1} = \left(\alpha_1 + A^{(1)} \right) \pm \sqrt{A^{(2)} - A^{(3)} \cos^2 y_1}$$
$$+ 2A^{(4)} \cos 2y_1 + 2A^{(5)} \cos \left(2y_1 + 2y_2 \right) + 2A^{(6)} \cos \left(2y_1 + 4y_2 \right) , \tag{31}$$

$$x_2 = \frac{\partial S}{\partial y_2} = \alpha_2 + 2A^{(5)} \cos\left(2y_1 + 2y_2\right) + 4A^{(6)} \cos\left(2y_1 + 4y_2\right),$$

$$x_3 = \frac{\partial S}{\partial y_3} = \alpha_3 ; \tag{32}$$

$$\frac{\partial S}{\partial \alpha_1} = n_1 t + \beta_1 = \left(1 + A_1^{(1)}\right) y_1 \pm M_1 \int \frac{dy_1}{\sqrt{A^{(2)} - A^{(3)} \cos^2 y_1}}$$

$$\pm \frac{1}{2} \frac{A_1^{(3)}}{A^{(3)}} \int \sqrt{A^{(2)} - A^{(3)} \cos^2 y_1}\, dy_1$$

$$+ A_1^{(4)} \sin 2y_1 + A_1^{(5)} \sin\left(2y_1 + 2y_2\right) + A_1^{(6)} \sin\left(2y_1 + 4y_2\right), \tag{33}$$

$$\frac{\partial S}{\partial \alpha_2} \quad n_2 t + \beta_2 = A_2^{(1)} y_1 + y_2 \pm M_2 \int \frac{dy_1}{\sqrt{A^{(2)} - A^{(3)} \cos^2 y_1}}$$

$$\pm \frac{1}{2} \frac{A_2^{(3)}}{A^{(3)}} \int \sqrt{A^{(2)} - A^{(3)} \cos^2 y_1}\, dy_1$$

$$+ A_2^{(4)} \sin 2y_1 + A_2^{(5)} \sin\left(2y_1 + 2y_2\right) + A_2^{(6)} \sin\left(2y_1 + 4y_2\right), \tag{34}$$

$$\frac{\partial S}{\partial \alpha_3} = n_3 t + \beta_3 = A_3^{(1)} y_1 + y_3 \pm M_3 \int \frac{dy_1}{\sqrt{A^{(2)} - A^{(3)} \cos^2 y_1}}$$

$$\pm \frac{1}{2} \frac{A_3^{(3)}}{A^{(3)}} \int \sqrt{A^{(2)} - A^{(3)} \cos^2 y_1}\, dy_1$$

$$+ A_3^{(4)} \sin 2y_1 + A_3^{(5)} \sin\left(2y_1 + 2y_2\right) + A_3^{(6)} \sin\left(2y_1 + 4y_2\right). \tag{35}$$

In this representation the terms containing y_2 in the argument are of the first order with respect to parameter k_2. In addition, these terms contain even powers of the eccentricity as a factor and consequently, will be small from the start; the computation of the second order term in k_2 with the argument y_2 shows that it can be neglected. All quantities can be considered as functions of $n_i t + \beta_i$, $(i = 1,2,3)$ but, of course, only y_2 and y_3 will have a secular term, since y_1 does not possess any such term.

It is not difficult to continue the process of computing S, if necessary. However, taking present day knowledge of the numerical values of geodetic parameters into consideration, it was found that even the development of equations 31 through 35 proved to be accurate—from the practical point of view—overly accurate. In this solution R_1 was originally considered to be of the second order with respect to w_{10}. It must be pointed out that this classification is purely formalistic, and loses its significance after the development is completed. The important characteristic of the solution is that it can be developed into a series in w, with the coefficients depending upon α_1, α_2, α_3. The development is not made in powers of $(k_2)^{1/2}$ as might be expected. This feature was observed initially by Izsak[3] in his solution of the critical inclination problem.

The method presented here does not introduce the small divisor A in the determination of ϕ_n which appears in the expression for S_n:

$$S_n = \phi_n(y_1) + \text{trigonometric terms in } y_1 \text{ and } y_2.$$

Every ϕ_n' will be a polynomial in A, if the constant of energy C is decomposed properly, to remove the poles with respect to A. This can be easily shown by applying the "from n to $n+1$ proof," since the equation for the determination of ϕ_n' has the form:

$$\phi_n' w_{20} A + P_n (\phi_1', \phi_2', \cdots, \phi_{n-1}'; \cos^2 y_1) = C_n - 1 ,$$

where P_n is a polynomial in ϕ_1', ϕ_2', $\cdots$, $\phi_{n-1}' \cos^2 y_1$. The elimination of $\cos^2 y_1$, by means of equation 17 and the proper determination of C_{n-1}, will lead to the representation of ϕ_n' in polynomial form, providing it has been shown that ϕ_1', ϕ_2', $\cdots$, ϕ'_{n-1} are polynomials in A. Eliminating higher powers of A in favor of $\cos^2 y_1$, we deduce that

$$\phi_n' = \alpha_n (\cos^2 y_1) + A\beta_n (\cos^2 y_1) ,$$

where α_n and β_n are polynomials in $\cos^2 y_1$ with polynomial coefficients in w. This result is similar to that obtained by Izsak[4] for the critical inclination problem. For the partial derivative of S_n with respect to y_2 we have

$$\frac{\partial S_n}{\partial y_2} = T_n^{(0)} (y_1, y_2) + A T_n^{(1)} (y_1, y_2) ,$$

where $T_n^{(0)}$ and $T_n^{(1)}$ are trigonometric polynomials in $2y_1$ and $2y_2$ with polynomial coefficients in w. When integrated, this has been found to be purely trigonometrical with respect to y_2.

In the stable case y_1 will have a long period oscillation about $\pi/2$ or $3\pi/2$. It is also of interest to know the approximate period of this deviation in longitude. From equations 23 through 29, we have

$$A_1^{(1)} = -1 - \frac{1}{3}\frac{w_{30}}{w_{20}} w + O(w^2) , \tag{36}$$

$$A_1^{(2)} = 2w + \left[\frac{w_{30}}{w_{20}} - 3\frac{Q_0'}{Q_0}\right] w^2 + O(w^3) , \tag{37}$$

$$\frac{A_1^{(3)}}{A^{(3)}} = O(w) , \tag{38}$$

$$M_1 = w + \frac{1}{2}\left[\frac{w_{30}}{w_{20}} - 3\frac{Q_0'}{Q_0}\right] w^2 + O(w^3) , \tag{39}$$

$$\frac{M_1}{\sqrt{A^{(2)}}} = 1 - \frac{Q_0'}{Q_0} w . \tag{40}$$

Neglecting small long-period terms, we can write, from equation 32, the expression for the period of

libration T:

$$n_1 T = 2\left(1 + A_1^{(1)}\right)\left[\arccos\left(\frac{1}{k}\right) - \arccos\left(-\frac{1}{k}\right)\right]$$

$$+ 2\frac{M_1}{\sqrt{A^{(2)}}}\int_{\cos^{-1}(-1/k)}^{\cos^{-1}(1/k)} \frac{dy_1}{\sqrt{1 + k^2\cos^2 y_1}}$$

$$+ \frac{A_1^{(3)}}{A^{(3)}}\sqrt{A^{(2)}}\int_{\cos^{-1}(-1/k)}^{\cos^{-1}(1/k)} \sqrt{1 - k^2\cos^2 y_1}\, dy_1 ,$$

where

$$k^2 = \frac{A^{(3)}}{A^{(2)}} = \frac{4Q_0}{w^2 w_{20}} .$$

Setting

$$k\cos y_1 = \sin u ,$$

we have

$$\int_{\cos^{-1}(-1/k)}^{\cos^{-1}(1/k)} \frac{dy_1}{\sqrt{1 - k^2\cos^2 y_1}} = -\frac{2}{k}\int_0^{\pi/2} \frac{du}{\sqrt{1 - \frac{1}{k^2}\sin^2 u}} = -\frac{2}{k}K\left(\frac{1}{k}\right) ,$$

and

$$\int_{\cos^{-1}(-1/k)}^{\cos^{-1}(1/k)} \sqrt{1 - k^2\cos^2 y_1}\, dy_1 = -\frac{2}{k}\int_0^{\pi/2} \frac{\cos^2 u\, du}{\sqrt{1 - \frac{1}{k^2}\sin^2 u}} = -\frac{2}{k}K\left(\frac{1}{k}\right) - 2k\left[E\left(\frac{1}{k}\right) - K\left(\frac{1}{k}\right)\right],$$

where $K(1/k)$ and $E(1/k)$ are the standard elliptic integrals of the first and second kinds having the modulus $1/k$. In terms of the new variables,

$$n_1 T = 2\left(1 + A_1^{(1)}\right)\left[2\arccos\left(\frac{1}{k}\right) - \pi\right] - \frac{4}{k}\left[\frac{M_1}{\sqrt{A^{(2)}}} + \frac{1}{2}\frac{A_1^{(3)}}{A^{(3)}}\sqrt{A^{(2)}}\right]K\left(\frac{1}{k}\right)$$

$$- 2k\sqrt{A^{(2)}}\frac{A_1^{(3)}}{A^{(3)}}\left[E\left(\frac{1}{k}\right) - K\left(\frac{1}{k}\right)\right] . \qquad (41)$$

Substituting equations 36 through 40 into equation 41 yields:

$$n_1 T = -\frac{2}{3}\frac{w_{30}}{w_{20}} w\left[2\cos^{-1}\left(\frac{1}{k}\right) - \pi\right] - \frac{4}{k}\left(1 - \frac{Q_0'}{Q_0}w\right)K\left(\frac{1}{k}\right) + O\left(\frac{w^2}{k}\right) .$$

Equation 14 and 23 give the expression for $C = C_0 + C_1$, from which

$$n_1 = \frac{\partial C}{\partial \alpha_1} = -w_{20} w\left[1 - \frac{1}{2w_{20}}\left(\frac{w_{30}}{w_{20}} + 3\frac{Q_0'}{Q_0}\right)w\right]$$

and

$$T = + \frac{2K\left(\frac{1}{k}\right)}{\sqrt{w_{20} Q_0}} + \frac{2}{3}\frac{w_{30}}{w_{20}^2}\left[2\cos^{-1}\left(\frac{1}{k}\right) - \pi\right] - \frac{4K\left(\frac{1}{k}\right)}{k\, w_{20}}\left[\left(1 - \frac{3}{2}\frac{1}{w_{20}}\right)\frac{Q_0'}{Q_0} - \frac{1}{2}\frac{w_{30}}{w_{20}^2}\right] + O\left(\frac{w}{k}\right) . \tag{42}$$

In order to find the period in the vicinity of the libration point, we expand equation 42 in powers of $1/k$. Since w approaches zero as the libration becomes smaller,

$$\frac{1}{k} = + \frac{w}{2}\sqrt{\frac{w_{20}}{Q_0}} \to 0 ;$$

and for small values of $1/k$,

$$\cos^{-1}\left(\frac{1}{k}\right) = \frac{\pi}{2} - \frac{1}{k} - \frac{1}{6}\frac{1}{k^3} + \cdots ,$$

and

$$K\left(\frac{1}{k}\right) = \frac{\pi}{2}\left(1 + \frac{1}{4}\frac{1}{k^2} + \cdots\right) .$$

Consequently we deduce that

$$T = \frac{\pi}{\sqrt{Q_0 w_{20}}}\left\{1 - w\left[\frac{2}{3\pi}\frac{w_{30}}{w_{20}} + \frac{Q_0'}{Q_0} - \frac{1}{2w_{20}}\left(\frac{w_{30}}{w_{20}} + 3\frac{Q_0'}{Q_0}\right)\right]\right\} + O\left(\frac{1}{k^2}\right) .$$

UNSTABLE MOTION

If the condition (equation 22) is not satisfied, the motion is unstable and the Hamiltonian function takes the form

$$F = R_0 + R_1 ,$$

with R_0 and R_1 defined by equations 11 and 12. Performing operations similar to the previous ones, we find that

$$S = \left(\alpha_1 + B^{(1)}\right) y_1 + \alpha_2 y_2 + \alpha_3 y_3 + \int \sqrt{B^{(2)} + B^{(3)} \sin^2 y_1}\, dy_1$$
$$+ B^{(4)} \sin 2y_1 + B^{(5)} \sin\left(2y_1 + 2y_2\right) + B^{(6)} \sin\left(2y_1 + 4y_2\right) ,$$

where

$$B^{(1)} = - w - \frac{2}{3} w^2 \frac{w_{30}}{w_{20}} - \left(\frac{1}{3}\frac{w_{30}}{w_{20}} - \frac{Q_0'}{Q_0}\right)\frac{Q_0}{w_{20}} , \tag{43}$$

$$B^{(2)} = w^2\left[1 + \frac{1}{2}\left(\frac{w_{30}}{w_{20}} - \frac{Q_0'}{Q_0}\right)w\right]^2 , \tag{44}$$

$$B^{(3)} = \frac{4Q_0}{w_{20}}\left[1 + \frac{1}{2}\left(\frac{w_{30}}{w_{20}} - \frac{Q_0'}{Q_0}\right)w\right]^2 > 0 , \tag{45}$$

$$B^{(4)} = \left(\frac{1}{6}\frac{w_{30}}{w_{20}} - \frac{1}{2}\frac{Q_0'}{Q_0}\right)\frac{Q_0}{w_{20}} ,$$

$$B^{(5)} = -\frac{Q_1}{2w_{01}} ,$$

and

$$B^{(6)} = -\frac{Q_2}{4w_{01}} .$$

Despite the similarity between some A's and B's, they are not identical since different values of R_0 are used in each case. The constant of energy has the same analytical form as in the stable case:

$$C = -{}'R_0(\alpha_1, \alpha_2) + \left(\frac{1}{6}\frac{w_{30}}{w_{20}} + \frac{1}{2}\frac{Q_0'}{Q_0}\right)w^3 . \tag{46}$$

Thus we have, as in the previous case,

$$x_1 = \alpha_1 + B^{(1)} + \sqrt{B^{(2)} + B^{(3)}\sin^2 y_1} + 2B^{(4)}\cos 2y_1 + 2B^{(5)}\cos(2y_1 + 2y_2) + 2B^{(6)}\cos(2y_1 + 4y_2) ,$$

$$x_2 = \alpha_2 + 2B^{(5)}\cos(2y_1 + 2y_2) + 4B^{(6)}\cos(2y_1 + 4y_2) ,$$

$$x_3 = \alpha_3 ,$$

and, putting

$$B_j^{(i)} = \frac{\partial B^{(i)}}{\partial \alpha_j} ,$$

$$\frac{\partial C}{\partial \alpha_i} = n_i ,$$

$$N_j = \frac{1}{2}\left(B_j^{(2)} - B^{(2)}\frac{B_j^{(3)}}{B^{(3)}}\right) ,$$

we have, designating by the additive constants of integration β_1, β_2, β_3:

$$n_1 t + \beta_1 = (1 + B_1^{(1)})y_1 + N_1\int\frac{dy_1}{\sqrt{B^{(2)} + B^{(3)}\sin^2 y_1}} + \frac{1}{2}\frac{B_1^{(3)}}{B^{(3)}}\int\sqrt{B^{(2)} + B^{(3)}\sin^2 y_1}\, dy_1$$

$$+ B_1^{(4)}\sin 2y_1 + B_1^{(5)}\sin(2y_1 + 2y_2) + B_1^{(6)}\sin(2y_1 + 4y_2) ; \tag{47}$$

$$n_2 t + \beta_2 = B_2^{(1)} y_1 + y_2 + N_2 \int \frac{dy_1}{\sqrt{B^{(2)} + B^{(3)} \sin^2 y_1}} + \frac{1}{2} \frac{B_2^{(3)}}{B^{(3)}} \int \sqrt{B^{(2)} + B^{(3)} \sin^2 y_1}\, dy_1$$

$$+ B_2^{(4)} \sin 2y_1 + B_2^{(5)} \sin (2y_1 + 2y_2) + B_2^{(6)} \sin (2y_1 + 4y_2) ;$$

$$n_3 t + \beta_3 = B_3^{(1)} y_1 + y_3 + N_3 \int \frac{dy_1}{\sqrt{B^{(2)} + B^{(3)} \sin^2 y_1}} + \frac{1}{2} \frac{B_3^{(3)}}{B^{(3)}} \int \sqrt{B^{(2)} + B^{(3)} \sin^2 y_1}\, dy_1$$

$$+ B_3^{(4)} \sin 2y_1 + B_3^{(5)} \sin (2y_1 + 2y_2) + B_3^{(6)} \sin (2y_1 + 4y_2) .$$

In the unstable case

$$B^{(2)} + B^{(3)} \sin^2 y_1 > 0 \tag{48}$$

and the square root of equation 47 (and its reciprocal) may be developed into a Fourier series in $2y_1$; the argument y_1 will possess a real secular term, which is absent in the stable case. The coefficient of y_1 in the right side of equation 46 is

$$P = 1 + B_1^{(1)} + \frac{2N_1}{\pi} \int_0^{\pi/2} \frac{dy_1}{\sqrt{B^{(2)} + B^{(3)} \sin^2 y_1}} + \frac{1}{\pi} \frac{B_1^{(3)}}{B^{(3)}} \int_0^{\pi/2} \sqrt{B^{(2)} + B^{(3)} \sin^2 y_1}\, dy_1 \tag{49}$$

Substituting in the integrand $y_1 = (\pi/2) - \phi$, and

$$k^2 = \frac{B^{(3)}}{B^{(2)} + B^{(3)}} ,$$

we reduce the integrals to the normal form and equation 49 becomes:

$$P = 1 + B_1^{(1)} + \frac{2KN_1}{\pi \sqrt{B^{(2)} + B^{(3)}}} + \frac{E}{\pi} \cdot \frac{B_1^{(3)}}{B^{(3)}} \sqrt{B^{(2)} + B^{(3)}} , \tag{50}$$

where

$$K(k) = \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - k^2 \sin^2 \phi}} ,$$

and

$$E(k) = \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2 \phi}\, d\phi .$$

Taking equations 43, 44, and 45 into account, we further deduce that

$$B_1^{(1)} = -1 - \frac{1}{3} w \frac{w_{30}}{w_{20}} + O(w^2) , \tag{51}$$

$$B_1^{(2)} = 2w + \left(\frac{w_{30}}{w_{20}} - 3 \frac{Q_0'}{Q_0} \right) w^2 + O(w^3) , \tag{52}$$

$$\frac{B_1^{(3)}}{B^{(3)}} = O(w) , \tag{53}$$

and

$$N_1 = w + \frac{1}{2}\left(\frac{w_{30}}{w_{20}} - 3\frac{Q_0'}{Q_0}\right)w^2 + O(w^3) \tag{54}$$

It follows from equations 50 through 54, neglecting the terms of the second order in w, that:

$$P = -\frac{1}{3}w\frac{w_{30}}{w_{20}} + \frac{2k'K}{\pi}\left(1 - \frac{Q_0'}{Q_0}\right), \tag{55}$$

where k' is the complementary modulus, $k^2+k'^2=1$, and

$$k' = \frac{w}{\sqrt{B^{(2)} + B^{(3)}}}\left[1 + \frac{1}{2}\left(\frac{w_{30}}{w_{20}} - \frac{Q_0'}{Q_0}\right)w\right].$$

From equation 46, neglecting terms of the third order, we have

$$n_1 = \frac{\partial C}{\partial \alpha_1} = -ww_{20} + \frac{1}{2}\left(\frac{w_{30}}{w_{20}} + 3\frac{Q_0'}{Q_0}\right)w^2 . \tag{56}$$

Using equations 54 and 55, we can compute the mean motion $\nu_1=n_1/P$ of the argument y_1 with an accuracy up to the terms of the second order in w. The value of ν_1 denotes the speed with which the satellite will depart from its original position over the earth's surface during the course of time.

CONCLUSION

The theory of motion of a 24-hour satellite under the influence of the ellipticity of earth's equator has been developed using a resonance theory. The expressions for the elements of motion can be represented in the form of a series with respect to the parameter w, which would be closely associated with the mean motion of the main critical argument $(n-n')t+\Omega+\omega$ in a nonresonance case. Canonical elements and the Hamilton–Jacobi partial differential equations were used to solve the problem. This method of solution was chosen because of its flexibility with respect to the form of the integration constants which are adjusted so as to remove small divisors from the solution.

The conditions are established for stable and unstable types of motion: no severe restrictions are imposed on the values of inclination or eccentricity. The formulas are developed to the point where a numerical development can be easily accomplished for any particular case. Using this method, an extension of Hori's critical inclination theory[5] can be easily obtained. The method described herein can also be applied to a more general case. If the ratio of mean motions is $n/n'=1/p$, and not $n/n'=1$ as in our case, the theory given here can be easily extended by using the set of canonical variables:

$$x_1 = \frac{\sqrt{\mu a}}{p}, \qquad y_1 = pl + \Omega + \omega - n't ,$$

$$x_2 = \sqrt{\mu a}\left(\frac{1}{p} - \sqrt{1-e^2}\right), \qquad y_2 = -\omega ,$$

$$x_3 = \sqrt{\mu a}\left(\frac{1}{p} - \sqrt{1-e^2}\cos i\right), \qquad y_3 = -\Omega + n't .$$

Comparing this theory with observations can help to better determine the coefficient of ellipticity of the earth's equator

REFERENCES

1. BROUWER, D., "Solution of the Problem of Artificial Satellite Theory without Drag," *Astronom. J.*, **64**(9): 378–397, November 1959.
2. BLITZER, L., and BOUGHTON, E. M., et al., "Effect of Ellipticity of the Equator on 24-Hour Nearly Circular Satellite Orbits," *J. Geophys. Res.*, **67**(1): 329–335, January 1962.
3. BOHLIN, K., "Uber eine Neue Annaherungsmetode in der Storungstheorie," *Bihang till Kongl. Svenska Vetenskaps-Akademiens Handlingar*, **14,** Afd. 1, Paper No. 5, 1888.
4. IZSAK, I. G., "On the Critical Inclination in Satellite Theory," Smithsonian Inst. Astrophys. Observ., Res. in Space Sci. Spec. Rept. No. 90, Mar. 14, 1962.
5. HORI, G., "The Motion of an Artificial Satellite in the Vicinity of the Critical Inclination," *Astronom. J.*, **65**(5): 291–300, June 1960.

ON THE GENERAL PLANETARY PERTURBATIONS IN RECTANGULAR COORDINATES

PETER MUSEN AND LLOYD CARPENTER

Goddard Space Flight Center

A new theory of general planetary perturbations in rectangular coordinates is developed. Expansion of the potential in terms of multipoles and application of operator calculus yield a direct and automatic way of forming the equations for the perturbations of any order. The perturbations in the position vector are decomposed along the instantaneous position and velocity vectors of the undisturbed motion and along the normal to the plane of the undisturbed motion. This decomposition leads to a direct method of integration and final formulas that are in a convenient form for programming.

NOTATIONS

k the Gaussian constant.

$\mathbf{r}$ the position vector of the disturbed body.

$r = |\mathbf{r}|$.

$\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \cdots$ the perturbations in $\mathbf{r}$.

m the mass of the disturbed body.

M the mass of the sun.

$a, e, n,$ the elements of the undisturbed orbit.

$a', e', n', \cdots$ the elements of the orbit of the disturbing body.

$h = 1/[a(1-e^2)]^{1/2}$

g the mean anomaly of the disturbed body.

g_0 the value of g at the epoch.

g' the mean anomaly of the disturbing body.

g_0' the value of g' at the epoch.

f the true undisturbed anomaly.

ϵ the eccentric undisturbed anomaly.

$\mathbf{R}$ the unit vector normal to undisturbed orbit plane.

$\nabla = \mathbf{i}(\partial/\partial x) + \mathbf{j}(\partial/\partial y) + \mathbf{k}(\partial/\partial z)$, the del operator.

$\mathbf{r}'$ the position vector of the disturbing body.

$r' = |\mathbf{r}'|$.

$\mathbf{r}_1', \mathbf{r}_2', \mathbf{r}_3', \cdots$ the perturbations in $\mathbf{r}'$.

m' the mass of the disturbing body.

η the fictitious eccentric anomaly, to be replaced by ϵ after the integration.

$\varphi = (n'/n)(\epsilon - g_0) + g_0'$

INTRODUCTION

In this article a new theory of general planetary perturbations in rectangular coordinates is developed. It is based on the decomposition of the potential in terms of multipoles with the axial vectors equal to the perturbations of different orders, in a way similar to that of Dziobek (1892). Another characteristic of this theory is that we have applied operator calculus, instead of using the method of variation of elements as an auxiliary means as was done in the earlier work (Brouwer, 1944).

The use of operator calculus is a more direct and automatic way of forming the differential equations for perturbations. The form of the variational equation to which the problem is reduced suggests a form of decomposition of perturbations similar to the decomposition of the increment of the position vector in the vectorial method of orbit correction (Eckert and Brouwer, 1937).

The perturbations in the position vector are decomposed along the instantaneous position and

velocity vectors of the undisturbed motion and along the normal to the plane of the undisturbed motion. Such a decomposition leads to a direct method of integration that is applicable to the perturbations of any order. The final developments are obtained in the form of trigonometric series with the number of arguments equal to the number of planets in the problem.

The problem of general perturbations in rectangular coordinates recently became the center of attention again. The works of Danby (1962), utilizing matrices, and of Gontkovskaja (1958), with the true anomaly as the independent variable, especially deserve to be mentioned. The general perturbations are definitely superior to the numerical integration procedure in the problems of orbital stability and long-term predictions.

Another reason for the revival of interest in general perturbations is the launching of artificial satellites and planets. The problem of motion of distant satellites in cislunar space is a typical planetary or cometary problem. The importance of artificial satellites for controlled experiments in celestial mechanics lies in the contraction of the time scale. Events that occur over a period of centuries in a planetary system take place in a few years for an artificial satellite under the influence of the moon. Finally the development of electronic computers has opened before the celestial mechanician possibilities that have never before existed in the history of science.

With modern computing machinery it is now convenient to compute general planetary perturbations for any cases of interest. Hansen's classical theory has been programmed for the IBM 7094 for automatic computation of first-order perturbations. About one minute of computer time is required for each planet. The symmetric form of the equations in the present theory facilitates the programing for perturbations of higher orders. First-order perturbations computed by this method have been compared with the results of Hansen's (1857–59) theory.

DEVELOPMENT OF BASIC EQUATIONS

The equation of the disturbed motion can be written in the form

$$d^2\mathbf{r}/dt^2=\mu_2\nabla(r^{-1}+\Omega) \tag{1}$$

where

$$\mu^2=k^2(M+m)$$

$$\Omega=\frac{m'}{M+m}\left(\frac{1}{|\mathbf{r}'-\mathbf{r}|}-\frac{\mathbf{r}\cdot\mathbf{r}'}{r'^3}\right) \tag{2}$$

The symbol ∇ is the del operator with respect to $\mathbf{r}$;

$$\nabla=\mathrm{grad}_{\mathbf{r}} \tag{3}$$

and, similarly,

$$\nabla'=\mathrm{grad}_{\mathbf{r}'} \tag{4}$$

Designating by $\mathbf{r}_0$ the undisturbed value of $\mathbf{r}$ and by $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \cdots$, the perturbations of the first, second, etc., orders, we have

$$\mathbf{r}=\mathbf{r}_0+\mathbf{r}_1+\mathbf{r}_2+\mathbf{r}_3+\cdots \tag{5}$$

and, similarly,

$$\mathbf{r}'=\mathbf{r}_0'+\mathbf{r}_1'+\mathbf{r}_2'+\mathbf{r}_3'+\cdots \tag{6}$$

Evidently

$$\mathrm{grad}_{\mathbf{r}}\varphi(\mathbf{r},\mathbf{r}')=\mathrm{grad}_{\mathbf{r}_0}\varphi(\mathbf{r}_0+\mathbf{r}_1+\mathbf{r}_2+\cdots,\mathbf{r}_0'+\mathbf{r}_1'+\cdots) \tag{7}$$

$$\mathrm{grad}_{\mathbf{r}'}\varphi(\mathbf{r},\mathbf{r}')=\mathrm{grad}_{\mathbf{r}_0'}\varphi(\mathbf{r}_0+\mathbf{r}_1+\cdots,\mathbf{r}_0'+\mathbf{r}_1'+\cdots) \tag{8}$$

In the further exposition we will discard the notations $\mathbf{r}_0$ and $\mathbf{r}_0'$ and will use the notations $\mathbf{r}$ and $\mathbf{r}'$ instead. In the same way we will use the symbols ∇ and ∇' to designate the del operators with respect to $\mathbf{r}_0$ and $\mathbf{r}_0'$.

Thus

$$\nabla=\mathrm{grad}_{\mathbf{r}}$$

$$\nabla'=\mathrm{grad}_{\mathbf{r}'}$$

as before, but a new meaning is attached to $\mathbf{r}$ and $\mathbf{r}'$. Equation 1 takes the form

$$(d^2/dt^2)(\mathbf{r}+\mathbf{r}_1+\mathbf{r}_2+\mathbf{r}_3+\cdots)=\mu^2\nabla\exp D(r^{-1}+\Omega) \tag{9}$$

where

$$D=(\mathbf{r}_1\cdot\nabla+\mathbf{r}_1'\cdot\nabla')+(\mathbf{r}_2\cdot\nabla+\mathbf{r}_2'\cdot\nabla')+(\mathbf{r}_3\cdot\nabla+\mathbf{r}_3'\cdot\nabla')+\cdots \tag{10}$$

The fact that $\mathbf{r}_i, \mathbf{r}_i'(i=1, 2, 3, \cdots)$ can be considered as constants in the process of applying the del operators leads to the development of

the force function in terms of multipoles with the axial vectors $\mathbf{r}_i$, $\mathbf{r}_i'(i=1, 2, 3, \cdots)$. Such a type of development is familiar from electrostatic theories. We have

$$\begin{aligned}\exp D = 1 &+ (\mathbf{r}_1\cdot\nabla+\mathbf{r}_1'\cdot\nabla') + [(\mathbf{r}_2\cdot\nabla+\mathbf{r}_2'\cdot\nabla') \\ &+ \tfrac{1}{2}(\mathbf{r}_1\cdot\nabla+\mathbf{r}_1'\cdot\nabla')^2] + [(\mathbf{r}_3\cdot\nabla+\mathbf{r}_3'\cdot\nabla') \\ &+ (\mathbf{r}_1\cdot\nabla+\mathbf{r}_1'\cdot\nabla')\cdot(\mathbf{r}_2\cdot\nabla+\mathbf{r}_2'\cdot\nabla') \\ &+ \tfrac{1}{6}(\mathbf{r}_1\cdot\nabla+\mathbf{r}_1'\cdot\nabla')^3] + \cdots \qquad (11)\end{aligned}$$

We use the following formulas representing the potentials of multipoles with the axial vectors $\mathbf{x}, \mathbf{y}, \mathbf{z}, \cdots$:

$$(\mathbf{x}\cdot\nabla)r^{-1} = -\mathbf{x}\cdot\mathbf{r}/r^3 \qquad (12)$$

$$(\mathbf{x}\cdot\nabla)(\mathbf{y}\cdot\nabla)r^{-1} = -\frac{\mathbf{x}\cdot\mathbf{y}}{r^3} + \frac{3\mathbf{x}\cdot\mathbf{r}\mathbf{y}\cdot\mathbf{r}}{r^5} \qquad (13)$$

$$\begin{aligned}(\mathbf{x}\cdot\nabla)(\mathbf{y}\cdot\nabla)(\mathbf{z}\cdot\nabla)r^{-1} &= (3/r^5)(\mathbf{r}\cdot\mathbf{x}\mathbf{y}\cdot\mathbf{z} \\ &+ \mathbf{r}\cdot\mathbf{y}\mathbf{z}\cdot\mathbf{x} + \mathbf{r}\cdot\mathbf{z}\mathbf{x}\cdot\mathbf{y}) - (15/r^7)\mathbf{x}\cdot\mathbf{r}\mathbf{y}\cdot\mathbf{r}\mathbf{z}\cdot\mathbf{r} \qquad (14)\end{aligned}$$

By applying (11) to r^{-1} and taking (12)–(14) into consideration, we have

$$\begin{aligned}(\exp D)r^{-1} = r^{-1} &- (\mathbf{r}_1\cdot\mathbf{r}/r^3) \\ &+ \left[-\frac{\mathbf{r}\cdot\mathbf{r}_2}{r^3} + \frac{3}{2}\frac{(\mathbf{r}\cdot\mathbf{r}_1)^2}{r^5} - \frac{1}{2}\frac{\mathbf{r}_1\cdot\mathbf{r}_1}{r^3}\right] \\ &+ \left[-\frac{\mathbf{r}\cdot\mathbf{r}_3}{r^3} + \frac{3\mathbf{r}\cdot\mathbf{r}_1\mathbf{r}\cdot\mathbf{r}_2}{r^5} - \frac{\mathbf{r}_1\cdot\mathbf{r}_2}{r^3}\right. \\ &\left. + \frac{3}{2}\frac{\mathbf{r}\cdot\mathbf{r}_1\mathbf{r}_1\cdot\mathbf{r}_1}{r^5} \quad \frac{5}{2}\frac{(\mathbf{r}\cdot\mathbf{r}_1)^3}{r^7}\right]\cdots \qquad (15)\end{aligned}$$

From the last equation we obtain

$$\begin{aligned}(\nabla\exp D)r^{-1} &= -\frac{\mathbf{r}}{r^3} + \left(-\frac{\mathbf{r}_1}{r^3} + \frac{3\mathbf{r}\mathbf{r}\cdot\mathbf{r}_1}{r^5}\right) \\ &+ \left[-\frac{\mathbf{r}_2}{r^3} + \frac{3\mathbf{r}\mathbf{r}\cdot\mathbf{r}_2}{r^5} + \frac{3\mathbf{r}\cdot\mathbf{r}_1\mathbf{r}_1}{r^5} + \frac{3}{2}\frac{\mathbf{r}\mathbf{r}_1\cdot\mathbf{r}_1}{r^5} - \frac{15}{2}\frac{\mathbf{r}(\mathbf{r}\cdot\mathbf{r}_1)^2}{r^7}\right] \\ &+ \left[-\frac{\mathbf{r}_3}{r^3} + \frac{3\mathbf{r}\mathbf{r}\cdot\mathbf{r}_3}{r^5} + \frac{3}{r^5}(\mathbf{r}\cdot\mathbf{r}_1\mathbf{r}_2 + \mathbf{r}\mathbf{r}_1\cdot\mathbf{r}_2\right. \\ &\left. + \mathbf{r}\cdot\mathbf{r}_2\mathbf{r}_1) - \frac{15}{r^7}\mathbf{r}\mathbf{r}_1\cdot\mathbf{r}\mathbf{r}_2\cdot\mathbf{r} + \frac{3}{2}\frac{\mathbf{r}_1\mathbf{r}_1\cdot\mathbf{r}_1}{r^5} - \frac{15}{2}\frac{\mathbf{r}\mathbf{r}\cdot\mathbf{r}_1\mathbf{r}_1\cdot\mathbf{r}_1}{r^7}\right. \\ &\left. - \frac{15}{2}\frac{\mathbf{r}_1(\mathbf{r}\cdot\mathbf{r}_1)^2}{r^7} + \frac{35}{2}\frac{\mathbf{r}(\mathbf{r}\cdot\mathbf{r}_1)^3}{r^9}\right]\cdots \qquad (16)\end{aligned}$$

Similarly we obtain for the disturbing force

$$\begin{aligned}(\nabla\exp D)\Omega = \nabla\Omega &+ (\mathbf{r}_1\cdot\nabla\nabla\Omega + \mathbf{r}_1'\cdot\nabla'\nabla\Omega) \\ &+ [(\mathbf{r}_2\cdot\nabla\nabla\Omega + \mathbf{r}_2'\cdot\nabla'\nabla\Omega) \\ &+ \tfrac{1}{2}(\mathbf{r}_1\cdot\nabla+\mathbf{r}_1'\cdot\nabla')^2\nabla\Omega] + \cdots \qquad (17)\end{aligned}$$

We have

$$\Omega = \frac{m'}{M+m}\left(\frac{1}{\rho} - \frac{\mathbf{r}'\cdot\mathbf{r}}{r'^3}\right) \qquad (18)$$

where

$$\boldsymbol{\rho} = \mathbf{r}' - \mathbf{r}$$

From (18) we deduce, besides the standard expression,

$$\nabla\Omega = \frac{m'}{M+m}\left(\frac{\boldsymbol{\rho}}{\rho^3} - \frac{\mathbf{r}'}{r'^3}\right), \qquad (19)$$

also the expressions

$$\nabla'\Omega = \frac{m'}{M+m}\left(-\frac{\boldsymbol{\rho}}{\rho^3} - \frac{\mathbf{r}}{r'^3} + \frac{3\mathbf{r}'\mathbf{r}'\cdot\mathbf{r}}{r'^5}\right) \qquad (20)$$

$$\nabla\nabla\Omega = \frac{m'}{M+m}\left(-\frac{I}{\rho^3} + \frac{3\boldsymbol{\rho}\boldsymbol{\rho}}{\rho^5}\right) \qquad (21)$$

$$\nabla'\nabla\Omega = \frac{m'}{M+m}\left(\frac{I}{\rho^3} - \frac{3\boldsymbol{\rho}\boldsymbol{\rho}}{\rho^5} - \frac{I}{r'^3} + \frac{3\mathbf{r}'\mathbf{r}'}{r'^5}\right) \qquad (22)$$

It follows from (21) – (22) that

$$\mathbf{r}_i\cdot\nabla\nabla\Omega = \frac{m'}{M+m}\left(-\frac{\mathbf{r}_i}{\rho^3} + \frac{3\boldsymbol{\rho}\boldsymbol{\rho}\cdot\mathbf{r}_i}{\rho^5}\right) \qquad (23)$$

$$\mathbf{r}_i'\cdot\nabla'\nabla\Omega = \frac{m'}{M+m}\left(\frac{\mathbf{r}_i'}{\rho^3} - \frac{3\boldsymbol{\rho}\boldsymbol{\rho}\cdot\mathbf{r}_i'}{\rho^5} - \frac{\mathbf{r}_i'}{r'^3} + \frac{3\mathbf{r}'\mathbf{r}'\cdot\mathbf{r}_i'}{r'^5}\right)$$
$$(i=1, 2, \cdots) \qquad (24)$$

Substituting (16) and (17) into (9) we deduce the differential equations for the computation of perturbations of different orders:

$$\frac{d^2\mathbf{r}}{dt^2} + \frac{\mu^2\mathbf{r}}{r^3} = 0 \quad \text{(the undisturbed motion)} \qquad (25)$$

$$\frac{d^2\mathbf{r}_1}{dt^2} + \mu^2\left(\frac{\mathbf{r}_1}{r^3} - \frac{3\mathbf{r}\mathbf{r}\cdot\mathbf{r}_1}{r^5}\right) = \mu^2\nabla\Omega \qquad (26)$$

$$\frac{d^2\mathbf{r}_2}{dt^2}+\mu^2\left(\frac{\mathbf{r}_2}{r^3}-\frac{3\mathbf{r}\mathbf{r}\cdot\mathbf{r}_2}{r^5}\right)=\mu^2(\mathbf{r}_1\cdot\nabla\nabla\Omega+\mathbf{r}_1'\cdot\nabla'\nabla\Omega)$$
$$+\frac{3\mu^2\mathbf{r}\cdot\mathbf{r}_1\mathbf{r}_1}{r^5}+\tfrac{3}{2}\mu^2\frac{\mathbf{r}\mathbf{r}_1\cdot\mathbf{r}_1}{r^5}-\tfrac{15}{2}\mu^2\frac{\mathbf{r}(\mathbf{r}\cdot\mathbf{r}_1)^2}{r^7} \quad (27)$$

$$\frac{d^2\mathbf{r}_3}{dt^2}+\mu^2\left(\frac{\mathbf{r}_3}{r^3}-\frac{3\mathbf{r}\mathbf{r}\cdot\mathbf{r}_3}{r^5}\right)=\mu^2(\mathbf{r}_2\cdot\nabla\nabla\Omega+\mathbf{r}_2'\cdot\nabla'\nabla\Omega)$$
$$+\tfrac{1}{2}\mu^2(\mathbf{r}_1\cdot\nabla+\mathbf{r}_1'\cdot\nabla')^2\nabla\Omega$$
$$+(3/r^5)\mu^2(\mathbf{r}\cdot\mathbf{r}_1\mathbf{r}_2+\mathbf{r}\mathbf{r}_1\cdot\mathbf{r}_2+\mathbf{r}\cdot\mathbf{r}_2\mathbf{r}_1)$$
$$-(15/r^7)\mu^2\mathbf{r}\mathbf{r}_1\cdot\mathbf{r}\mathbf{r}_2\cdot\mathbf{r}+\tfrac{3}{2}\mu^2(1/r^5)\mathbf{r}_1\mathbf{r}_1\cdot\mathbf{r}_1$$
$$-\tfrac{15}{2}\mu^2(1/r^7)\mathbf{r}\mathbf{r}\cdot\mathbf{r}_1\mathbf{r}_1\cdot\mathbf{r}_1$$
$$-\tfrac{15}{2}\mu^2(1/r^7)\mathbf{r}_1(\mathbf{r}\cdot\mathbf{r}_1)^2$$
$$+\tfrac{35}{2}\mu^2(1/r^9)\mathbf{r}(\mathbf{r}\cdot\mathbf{r}_1)^3 \quad (28)$$

All the equations defining the perturbations of different orders have the form of the variational equation

$$\frac{d^2\mathbf{x}}{dt^2}+\mu^2\left(\frac{I}{r^3}-\frac{3\mathbf{r}\mathbf{r}}{r^5}\right)\cdot\mathbf{x}=\mu^2\mathbf{F} \quad (29)$$

One of the most convenient substitutions to integrate this equation is the substitution

$$\mathbf{x}=T\mathbf{r}+(S/\mu)(d\mathbf{r}/dt)+z\mathbf{R} \quad (30)$$

A substitution of a similar type was used by Kulikov (1950) to develop the equations of the Eckert–Brouwer orbit correction method if $\mathbf{F}=0$. It can be used profitably also in the theory of general perturbations. In practical applications it is more convenient to put

$$d\mathbf{r}/dt=\mu\mathbf{v} \quad (31)$$

and to write (30) in the form

$$\mathbf{x}=T\mathbf{r}+S\mathbf{v}+z\mathbf{R} \quad (30')$$

We have

$$d\mathbf{v}/dt=-\mu\mathbf{r}/r^3 \quad (32)$$

We deduce from (30′)

$$\frac{d\mathbf{x}}{dt}=\left(\frac{dT}{dt}-\frac{\mu S}{r^3}\right)\mathbf{r}+\left(\mu T+\frac{dS}{dt}\right)\mathbf{v}+\frac{dz}{dt}\mathbf{R} \quad (30'')$$

Substituting (30′) into (29) and taking (32) into account we obtain

$$\left(\frac{d^2T}{dt^2}-\frac{3\mu^2}{r^3}T-\frac{2\mu}{r^3}\frac{dS}{dt}\right)\mathbf{r}+\left(\frac{d^2S}{dt^2}+2\mu\frac{dT}{dt}\right)\mathbf{v}$$
$$+\left(\frac{d^2z}{dt^2}+\frac{\mu^2z}{r^3}\right)\mathbf{R}=\mu^2\mathbf{F} \quad (33)$$

Multiplying (33) by

$$\mathbf{v}\times\mathbf{R} \quad \mathbf{R}\times\mathbf{r} \quad \text{and} \quad \mathbf{R}$$

and taking the area integral

$$h\mathbf{R}\cdot\mathbf{r}\times\mathbf{v}=1$$

into account, we obtain

$$\frac{d^2T}{dt^2}-\frac{3\mu^2}{r^3}T-\frac{2\mu}{r^3}\frac{dS}{dt}=\mu^2h\mathbf{F}\cdot\mathbf{v}\times\mathbf{R} \quad (34)$$

$$(d^2S/dt^2)+2\mu(dT/dt)=\mu^2h\mathbf{F}\cdot\mathbf{R}\times\mathbf{r} \quad (35)$$

Equations 34 and 35 represent the generalized Kulikov's results. In addition we have an equation of the standard form

$$(d^2z/dt^2)+(\mu^2z/r^3)=\mu^2\mathbf{F}\cdot\mathbf{R} \quad (36)$$

Equation 35 admits the first integral

$$\frac{dS}{dt}+2\mu T=\mu K_3+\int\mu^2h\mathbf{F}\cdot\mathbf{R}\times\mathbf{r}\,dt \quad (37)$$

Substituting (37) into (34) we obtain

$$\frac{d^2T}{dt^2}+\frac{\mu^2T}{r^3}=\mu^2h\mathbf{F}\cdot\mathbf{v}\times\mathbf{R}+\frac{2\mu^2K_3}{r^3}$$
$$+\frac{2}{r^3}\int\mu^3h\mathbf{F}\cdot\mathbf{R}\times\mathbf{r}\,dt \quad (38)$$

This equation can be integrated using Hill's (1874) procedure and the result is

$$T=K_1r\cos f+K_2r\sin f+\int\left(h^2\mu\mathbf{F}\cdot\mathbf{v}\times\mathbf{R}+\frac{2\mu hK_3}{r_3}\right.$$
$$\left.+\frac{2}{r^3}\int\mu^2h^2\mathbf{F}\cdot\mathbf{R}\times\mathbf{r}\,dt\right)\bar{r}r\sin(\bar{f}-f)\,dt \quad (39)$$

where $\bar{r}$ and $\bar{f}$ are considered as temporary constants; after the integration they are replaced by r and f.

Taking

$$\mathbf{R}\cdot\mathbf{r}\times\bar{\mathbf{r}}=\bar{r}r\sin(\bar{f}-f)$$

into account we can write (39) in the form

$$T=K_1r\cos f+K_2r\sin f+\int\left(h^2\mu\mathbf{F}\cdot\mathbf{v}\times\mathbf{R}+\frac{2h\mu K_3}{r^3}\right.$$

$$\left.+\frac{2}{r^3}\int h^2\mu^2\mathbf{F}\cdot\mathbf{R}\times\mathbf{r}\,dt\right)\mathbf{R}\cdot\mathbf{r}\times\bar{\mathbf{r}}\,dt \quad (40)$$

Considering the fact that $\bar{\mathbf{r}}$ is a relative constant, and taking (32) into account, we have

$$(\mu/r^3)\mathbf{R}\cdot\mathbf{r}\times\bar{\mathbf{r}}=(d/dt)\mathbf{R}\cdot\bar{\mathbf{r}}\times\mathbf{v} \quad (41)$$

Integrating (40) by parts and taking (32) and (41) into account we obtain

$$T=K_1r\cos f+K_2r\sin f+2K_3+A \quad (42)$$

where

$$A=\int h^2\mu(\mathbf{F}\cdot\mathbf{v}\times\mathbf{R})(\mathbf{R}\cdot\mathbf{r}\times\bar{\mathbf{r}})\,dt$$

$$+2\int h\mu(\mathbf{F}\cdot\mathbf{R}\times\mathbf{r})(1+h\mathbf{R}\cdot\mathbf{v}\times\bar{\mathbf{r}})\,dt \quad (43)$$

Taking

$$\int r\cos f\,dt=\frac{a}{n}[-\tfrac{3}{2}ent$$

$$+(1-\tfrac{1}{2}e^2)\sin\epsilon-\tfrac{1}{4}e\sin 2\epsilon] \quad (44)$$

$$\int r\sin f\,dt=\frac{b}{n}(-\cos\epsilon+\tfrac{1}{4}e\cos 2\epsilon) \quad (45)$$

into account we deduce from (37) and (42)

$$S=+2a^{5/2}K_1[\tfrac{3}{2}ent+(\tfrac{1}{2}e^2-1)\sin\epsilon+\tfrac{1}{4}e\sin 2\epsilon]$$

$$+2a^{5/2}(1-e^2)^{1/2}K_2\cdot(+\cos\epsilon-\tfrac{1}{4}e\cos 2\epsilon)$$

$$-3\mu K_3t+K_4+B \quad (46)$$

where

$$B=\iint\mu^2h\mathbf{F}\cdot\mathbf{R}\times\mathbf{r}\,dt^2-2\int\mu A\,dt \quad (47)$$

Similarly we obtain from (36)

$$z=K_5r\cos f+K_6r\sin f+C \quad (48)$$

where

$$C=\int h\mu(\mathbf{F}\cdot\mathbf{R})(\mathbf{R}\cdot\mathbf{r}\times\bar{\mathbf{r}})\,dt \quad (49)$$

CONSTANTS OF INTEGRATION

The determination of the constants of integration is especially simple if the initial conditions for $t=0$ are given.

In the further exposition we will use the notations

$$\varphi_0=\varphi(t)|_{t=0}$$

$$\dot{\varphi}_0=d\varphi/dt|_{t=0}$$

$$\varphi_0'=d\varphi/d\epsilon|_{t=0}$$

Let us assume that T_0, S_0, $\dot{T}_0$, $\dot{S}_0$ are given and we have to determine the constants of integration. From (37) we deduce immediately

$$K_3=+2T_0+(1/\mu)\dot{S}_0-\left(\int\mu h\mathbf{F}\cdot\mathbf{R}\times\mathbf{r}\,dt\right)_{t=0} \quad (50)$$

From (42) we obtain

$$K_1r_0\cos f_0+K_2r_0\sin f_0=T_0-2K_3-A_0 \quad (51)$$

and after differentiating (42) we have

$$-K_1\mu h\sin f_0+K_2\mu h\cdot(\cos f_0+e_0)=\dot{T}_0-\dot{A}_0 \quad (52)$$

From (51) and (52) we have

$$K_1=(T_0-A_0-2K_3)h^2(\cos f_0+e)$$

$$-(h/\mu)(\dot{T}_0-\dot{A}_0)r_0\sin f_0 \quad (53)$$

$$K_2=(T_0-A_0-2K_3)h^2\sin f_0$$

$$+(h/\mu)(\dot{T}_0-\dot{A}_0)r_0\cos f_0 \quad (54)$$

Taking the formula

$$\frac{d\varphi}{dt}=\frac{\mu}{r(a)^{1/2}}\frac{d\varphi}{d\epsilon}$$

into account, we can write (53) and (54) in the form

$$K_1=(T_0-A_0-2K_3)\frac{\cos f_0+e}{a(1-e^2)}$$

$$-(T_0'-A_0')\frac{\sin f_0}{a(1-e^2)^{1/2}} \quad (55)$$

$$K_2=(T_0-A_0-2K_3)\frac{\sin f_0}{a(1-e^2)}$$

$$+(T_0'-A_0')\frac{\cos f_0}{a(1-e^2)^{1/2}} \quad (56)$$

and similarly we have

$$K_5 = (z_0 - C_0)\frac{\cos f_0 + e}{a(1-e^2)} - (z_0' - C_0')\frac{\sin f_0}{a(1-e^2)^{1/2}} \quad (57)$$

$$K_6 = (z_0 - C_0)\frac{\sin f_0}{a(1-e^2)} + (z_0' - C_0')\frac{\cos f_0}{a(1-e^2)^{1/2}} \quad (58)$$

Finally, from (46) we obtain

$$K_4 = (S_0 - B_0) + 2K_1\mu\left(\int r\cos f\,dt\right)_{t=0} + 2K_2\mu\left(\int r\sin f\,dt\right)_{t=0} \quad (59)$$

where the integrals

$$\int r\cos f\,dt \quad \text{and} \quad \int r\sin f\,dt$$

are taken in the form given by (44) and (45), putting $t=0$.

If the undisturbed elements are osculating at the moment $t=0$, then we put

$$T_0 = \dot{T}_0 = S_0 = \dot{S}_0 = z_0 = \dot{z}_0 = 0$$

There exists also the set of elements such that

$$T_0 = A_0 \qquad \dot{T}_0 = \dot{A}_0$$
$$S_0 = B_0 \qquad \dot{S}_0 = \dot{B}_0$$
$$z_0 = C_0 \qquad \dot{z}_0 = \dot{C}_0$$

and consequently

$$K_1 = K_2 = K_3 = K_4 = K_5 = K_6 = 0$$

In other words, the integration in (42), (46), and (49) is performed only formally, without adding any constant. Thus

$$T = A$$
$$S = B$$
$$z = C$$

where A, B, and C are given by (43), (47), and (49). This set of elements is normally used in many text books of celestial mechanics in the representation of planetary theories. Of course, its determination is associated with the repetition of the orbit correction and the repetition of computation of general perturbations, until the convergence is reached.

With the program available at the Theoretical Division of the Goddard Space Flight Center such a procedure does not present any difficulties.

COLLECTION OF FORMULAS FOR THE PLANETARY CASE AND PROGRAMMING

For the computation of the perturbations we start with (42) through (49). The method of computation will vary slightly, depending on whether the integration is done with respect to the time or the eccentric anomaly. When we use the eccentric anomaly, the expressions other than **F** take the form of simple polynomials in the cosines and sines of the angles ϵ and η. In terms of the mean anomaly these are replaced by series. The components of the disturbing force are expanded into trigonometric series by means of numerical double harmonic analysis.

The amount of computation is reduced if the coordinate axes are taken in the directions of **P**, **Q**, and **R**. The matrix {**P'Q'R'**} is transformed to this reference by premultiplying by the transpose of {**PQR**}.

A large part of the computation will be spent in the expansion of the components of the disturbing force. The problem is reduced to double harmonic analysis of the odd powers of $1/\rho$, where ρ is the mutual distance. Hansen's method using Laplace coefficients has been programmed by one of us and applied with complete success to the computation of Hansen's perturbations of minor planets. However, for highly eccentric orbits such as that of Encke's comet the iteration process involved in this method does not converge rapidly. A program for direct double harmonic analysis has been developed by Neal Paris of the Theoretical Division. The initial results from this program indicate that this method is the most efficient and widely applicable. The development can be done in the mean anomalies, the eccentric anomalies, or in the variables ϵ and φ so that the Bessel transformations are eliminated. A collection of formulas will be given for integration in terms of the eccentric anomaly. This will also facilitate the comparison with Hansen's classical method as given in the Auseinandersetzung.

For the double harmonic analysis the numerical

values of the components of the disturbing force corresponding to given values of ϵ and φ are obtained from the equations

$$r = a(1 - e\cos\epsilon)$$

$$\mathbf{r} = a\mathbf{P}(\cos\epsilon - e) + b\mathbf{Q}\sin\epsilon$$

$$\epsilon' - e'\sin\epsilon' = \varphi - (n'/n)e\sin\epsilon$$

$$r' = a'(1 - e'\cos\epsilon')$$

$$\mathbf{r}' = a'\mathbf{P}'(\cos\epsilon' - e') + b'\mathbf{Q}'\sin\epsilon'$$

and

$$\rho^2 = r^2 + r'^2 - 2\mathbf{r}\cdot\mathbf{r}'$$

For the first-order perturbations we have

$$\mathbf{F} = m'\left(\frac{\mathbf{r}' - \mathbf{r}}{\rho^3} - \frac{\mathbf{r}'}{r'^3}\right)$$

If we put

$$N_1 = \sin(\eta - \epsilon) - \epsilon\sin\eta + e\sin\epsilon$$

$$N_2 = 2 - 2\cos(\eta - \epsilon)$$

$$N_3 = 1 - e\cos\epsilon$$

$$M_1 = a^2\mathbf{F}\cdot[\mathbf{P}\cos\epsilon + (1-e^2)^{-1/2}\mathbf{Q}\sin\epsilon]$$

$$M_2 = a^2\mathbf{F}\cdot[-\mathbf{P}\sin\epsilon + (1-e^2)^{-1/2}\mathbf{Q}(\cos\epsilon - e)]$$

$$M_3 = a^3\mathbf{F}\cdot\mathbf{R}(1 - e\cos\epsilon)$$

and

$$M_4 = \int M_2 N_3\, d\epsilon$$

we have

$$A = \int (M_1 N_1 + M_2 N_2)\, d\epsilon$$

$$B = a^{3/2}\int (M_4 - 2A) N_3\, d\epsilon$$

and

$$C = \int M_3 N_1\, d\epsilon$$

The value of M_4 at $t=0$ will be needed in the determination of the constants of integration. It is understood that η is replaced by ϵ after each integration.

For computing the constants of integration the series A, B, C and their derivatives are evaluated at $t=0$. From these values and the initial conditions we have

$$K_3 = 2T_0 + [1/r_0(a)^{1/2}]S_0' - (M_4)_0$$

$$K_1 = (T_0 - A_0 - 2K_3)\frac{\cos\epsilon_0}{r_0} - (T_0' - A_0')\frac{\sin\epsilon_0}{r_0}$$

$$K_2 = (T_0 - A_0 - 2K_3)\frac{\sin\epsilon_0}{r_0(1-e^2)^{1/2}} + (T_0' - A_0')\frac{\cos\epsilon_0 - e}{r_0(1-e^2)^{1/2}}$$

$$K_5 = (z_0 - C_0)\frac{\cos\epsilon_0}{r_0} - (z_0' - C_0')\frac{\sin\epsilon_0}{r_0}$$

$$K_6 = (z_0 - C_0)\frac{\sin\epsilon_0}{r_0(1-e^2)^{1/2}} + (z_0' - C_0')\frac{\cos\epsilon_0 - e}{r_0(1-e^2)^{1/2}}$$

and

$$K_4 = (S_0 - B_0) + a^{5/2}[(2-e^2)\sin\epsilon_0 - \tfrac{1}{2}e\sin 2\epsilon_0]K_1 + a^{5/2}(1-e^2)^{1/2}[-2\cos\epsilon_0 + \tfrac{1}{2}e\cos 2\epsilon_0]K_2$$

Adding the constants of integration to the series we have

$$T = A + (2K_3 - aeK_1) + aK_1\cos\epsilon + bK_2\sin\epsilon$$

$$S = B + K_4 + 3a^{3/2}(aeK_1 - K_3)(nt) + 2a^{5/2}(1-e^2)^{1/2}K_2\cos\epsilon - 2a^{5/2}(1 - \tfrac{1}{2}e^2)K_1\sin\epsilon$$

$$\tfrac{1}{2}[a^{5/2}e(1-e^2)^{1/2}]K_2\cos 2\epsilon + \tfrac{1}{2}(a^{5/2}e)K_1\sin 2\epsilon$$

and

$$z = C - aeK_5 + aK_5\cos\epsilon + bK_6\sin\epsilon$$

Finally, we have for the perturbations

$$\delta\mathbf{r} = T\mathbf{r} + S\mathbf{v} + z\mathbf{R}$$

where $\delta\mathbf{r} = \mathbf{r}_1, \mathbf{r}_2, \cdots$, and T, S, and z are different for the perturbations of different orders. The components of $\delta\mathbf{r}$ are referred to the rectangular coordinates by taking

$$\mathbf{r} = a\mathbf{P}(\cos\epsilon - e) + b\mathbf{Q}\sin\epsilon$$

and

$$\mathbf{v} = [1/r(a)^{1/2}][-a\mathbf{P}\sin\epsilon + b\mathbf{Q}\cos\epsilon]$$

with $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{R}$ now referred to the equator in the usual manner.

CONCLUSION

The system described in this article shares certain common characteristics with Hill's method. In our system we have the decomposition of perturbations along **r**, **v**, and **R** and in Hill's method along **r**, **R**×**r**, **R**. The directions **v** and **R**×**r** do not differ much in the case of a planetary motion. We integrate the equations for T and z using the same procedure as in Hill's method.

Hill's method introduces a double and a triple integral and seven interdependent constants of integration. In the presented method there are six constants and they are determined in a direct manner.

The theory of general perturbations in rectangular coordinates has also all positive characteristics of Hansen's planetary theory. It is purely numerical, and it is valid also for circular orbit.

The computation of perturbations of higher orders is simpler in the method of rectangular coordinates than in Hansen's method.

The decomposition of the potential in terms of multipoles, the application of operator calculus, and the substitution (30) is a direct and homogeneous way to obtain the perturbations of any order in rectangular coordinates. Thus the application of variation of elements, being unessential to the problem, is elimina[illegible]

The final formulas are brought to the form that facilitates the programing up to perturbations of third order.

REFERENCES

BROUWER, D., Integration of the equations of general planetary theory in rectangular coordinates, *Astron. J.*, **51,** 37–43, 1944.

DANBY, J. M. A., Integration of the equations of planetary motion in rectangular coordinates, *Astron. J.*, **67,** 287–299, 1962.

DZIOBEK, O., *Mathematical Theories of Planetary Motions,* Register Publishing Co., 1892. (Republished by Dover, New York, 1962.)

ECKERT, W. J., and BROUWER, D., The use of rcetangular coordinates in the differential correction of orbits, *Astron. J.*, **46,** 125, 1937.

GONTKOVSKAJA, V. T., The application of modern computational techniques to the analytical methods of celestial mechanics (Russian), *Bull. Inst. Theoret. Astron.*, **82,** 592–629, 1958.

HANSEN, D. A., *Auseinandersetzung einer zweckmassiger methode,* Leipzig, 1857–59.

HILL, G. W., A method of computing absolute perturbations, *Aston. Nachr.*, **83,** 209, 1874.

KULIKOV, D. K., Numerical methods of celestial mechanics as applied to the investigation of the satellite Jupiter 8 (Russian), *Bull. Inst. Theoret. Astron.*, **60,** 313–340, 1950.

STRESS DIFFERENCES AND THE REFERENCE ELLIPSOID

JOHN A. O'KEEFE AND WILLIAM M. KAULA

Goddard Space Flight Center

In a recent communication, Hulley[1] has connected gravity anomalies with other geophysical phenomena including faults and the pole positions. Unfortunately, the latter suggestion is not substantiated mathematically; for areas of any extent and for realistic rheology the *polfluchtkraft* can even be in the direction opposite that shown in Hulley's diagram.[2] In this work, Hulley made use of diagrams of the contours of the geoid supplied by Kaula.[3] The geoid contours to which Hulley refers do not give a clear picture of the distribution of the stress differences. This is because the reference ellipsoid is an approximation to the average ellipsoid. Stress differences, however, arise from the difference between the actual form of the earth and the theoretical one for fluid equilibrium. The flattening which corresponds to fluid equilibrium is approximately 1/300 as was pointed out by Henriksen[4] and later discussed by O'Keefe[5] and Munk and MacDonald.[6] If we plot the values of the gravity anomalies referred to an ellipsoid with a flattening of 1/300, we get the result as shown in fig. 1, which is based on Kaula's work. In comparison with Hulley's paper, fig. 1 indicates that there may be a relation between the tectonic activity and gravity anomalies: at least

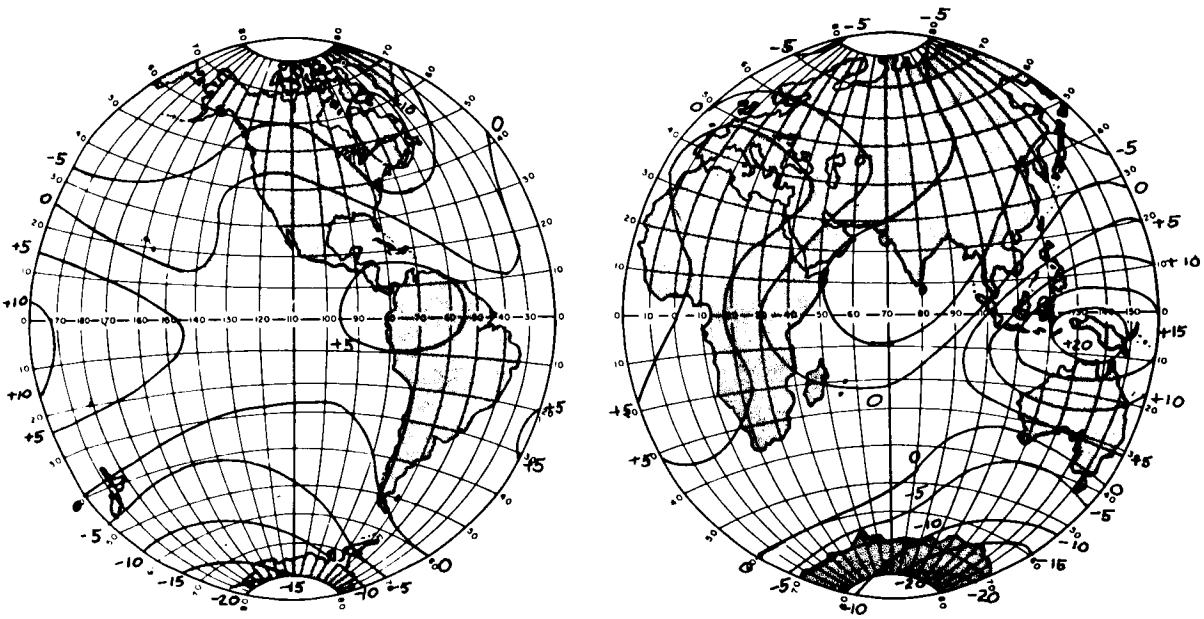


FIGURE 1.—Gravity anomalies, in milligals, derived from satellite perturbations and referred to an ellipsoid with a flattening of 1/299.8.

the strong positive anomalies in the East Indian area appear to correspond with the maximum tectonic activity.

On the other hand, it should also be pointed out that there is a special explanation associated with the largest part of the discrepancy between the actual and equilibrium figures: the difference in oblateness can be considered as a lag of 10^7 years in adjustment to the slowing of the earth's rotation.[6] So it is not entirely clear what the proper reference figure should be.

It is interesting to note that, regardless of the reference figure used, the shape of the geoid does not lend any particular support to the suggestion of Girdler[7] that the rift valleys and the mid-ocean ridges are the loci of up-currents in a convection system. It has been shown, by Licht[8] for example, that the top of a convection current should be in the area of positive gravity anomalies.

The positive anomaly areas near Central America, West Africa, and the East Indies are not associated with any ocean ridges. On the contrary, the ocean ridge system extending from the Northwest Indian Ocean, around south of Australia, and up to the East Pacific is strongly correlated with a negative belt in the gravity field.

A similar negative correlation exists between heat flow and the gravity field, as shown by Lee and MacDonald,[9] whose harmonic analysis of thermal measurements shows areas of maximum heat flow in central Asia and the Eastern Pacific, and areas of minimum heat flow in the South Atlantic and Western Pacific.

The various correlations shown are suggestive of what hypotheses to pursue, but they undoubtedly have a strong subjective element, and need both firmer mathematical models and more extensive data: in particular, more widespread gravimetry.

REFERENCES AND NOTES

1. HULLEY, J. L. C., *Nature*, **198**, 466 (1963).
2. SCHEIDEGGER, A. H., *Principles of Geogynamics* (Springer, Berlin, ed. 2, 1963), p. 170.
3. KAULA, W. M., *J. Geophys. Res.*, **64**, 2401 (1959); **66**, 1799 (1961); **68**, 473 (1963).
4. HENRIKSEN, S. W., *Ann. Intern. Geophys. Yr.*, **12**, 197 (1960).
5. O'KEEFE, J. A., *J. Geophys. Res.*, **64**, 2389 (1959).
6. MUNK, W. H., and MACDONALD, G. J. F., *ibid.*, **65**, 2169 (1960).
7. GIRDLER, R. W., *Nature*, **198**, 1037 (1963).
8. LICHT, A. L., *J. Geophys. Res.*, **65**, 349 (1960).
9. LEE, W. H. C., and MACDONALD, G. J. F., *ibid.*, in press.

ON LAGRANGE'S THEORY OF THE THREE-BODY PROBLEM

KARL STUMPFF*

Goddard Space Flight Center

Lagrange's famous studies of the three-body problem, presented in Tisserand's *Traité de Mécanique Céleste*, vol. I, are redeveloped in modern mathematical expressions. This new portrayal of a very old problem of celestial mechanics gives rise to various considerations which can be very useful for the practical computation. In particular, the nine elements of the "reduced" three-body problem are defined as a symmetrically constructed system of three groups of three elements each, and their differential equations are shown. A byproduct of this study is a representation of the coefficients of the equation of the fourth degree for Lagrange's quantity ρ in symmetrical form.

INTRODUCTION

Lagrange's famous treatment of the three-body problem is still the point of departure of many studies on this difficult problem, which has never been solved in its totality. Although Lagrange's theory is interpretated excellently in its mathematical elegance and clarity in Tisserand's textbook of celestial mechanics,** it is intriguing and, for modern application, practical to bring forth even more the insight and beauty of this theory by using modern mathematical representation. The following considerations are intended mainly to serve the end of bringing this classic theory, in new trappings, into the focus of celestial mechanical study. Also, in the reconsideration of this old problem, certain previously ignored questions will be answered.

Lagrange presented the following postulate: The general three-body problem can be solved completely if it is possible to represent as functions of time all those factors which are independent of the special selection of the coordinate system. The *relative* three-body problem (i.e., the problem of the motion of two bodies in reference to a third or of all three in reference to the common center of gravity) requires 12 integrals, 3 of which determine the orientation of the system in space. Thus by Lagrange's theorem only nine integrals must be found, for the other three can then be determined easily by a simple quadrature based upon functions known.

According to Hesse, we call this problem of seeking the nine "geometric" integrals the "reduced" three-body problem.*** It is easy to see that this reduced problem can actually be satisfied by nine quantities which are invariable with respect to coordinate transformation. The geometric figure formed by the four vectors determining motion—the vectors of location and velocity of the two bodies with reference to the third—is determined by nine quantities which are independent of the coordinate system. These nine quantities are the values of these four vectors and the five angles by which the vectors' respective locations are fixed. The construction of this figure from these nine elements is not significant in itself, since the components can be assembled in several different ways; but of the various possibilities there will always be only one that is compatible with the initial conditions given.

These facts, easily understandable geomet-

*NAS-NASA Research Associate; Professor Emeritus, Göttingen University.

**Tisserand, F., "*Traité de Mécanique Céleste*," Paris: Gauthier-Villars et Fils, 1889, vol. I.

***This is not to be confused with the *restricted* three-body problem.

tions of equations 4 and 5. For instance,

$$p_{15}=(\mathbf{r}_1\mathbf{r}_5)=(\mathbf{r}_1,\ -\mathbf{r}_1-\mathbf{r}_3)=-p_{11}-p_{13}, \quad (6)$$

$$p_{55}=(\mathbf{r}_5\mathbf{r}_5)=(-\mathbf{r}_1-\mathbf{r}_3)^2=p_{11}+p_{33}+2p_{13}. \quad (7)$$

The second of these equations is identical to the cosine theorem of plane trigonometry applied to the triangle of the three bodies in figures 1a and 1b.

Equation 3 is not symmetrical since it contains only the invariables with the indexes 1 to 4; it can be made symmetrical, however, if we use Lagrange's quantity ρ.

Lagrange noted that the remarkable relationship

$$2\rho=p_{14}-p_{23}=p_{25}-p_{16}=p_{36}-p_{45} \quad (8)$$

exists among the invariables, for

$$p_{14}-p_{23}=-(p_{16}+p_{12})+(p_{25}+p_{21})=p_{25}-p_{16},$$

and

$$p_{14}-p_{23}=-(p_{34}+p_{54})+(p_{43}+p_{63})=p_{36}-p_{45},$$

On the other hand, if we set

$$\left.\begin{aligned} 2s_1 &= p_{36}+p_{45},\\ 2s_3 &= p_{25}+p_{16},\\ 2s_5 &= p_{14}+p_{23}, \end{aligned}\right\} \quad (9)$$

from equations 7 and 8 we obtain

$$\begin{aligned} p_{14} &= s_5+\rho, \quad p_{25}=s_3+\rho, \quad p_{36}=s_1+\rho\\ p_{23} &= s_5-\rho, \quad p_{16}=s_3-\rho, \quad p_{45}=s_1-\rho. \end{aligned} \quad (10)$$

If equation 10 is substituted into equation 3 for p_{14} and p_{23},

$$M=\begin{vmatrix} p_{11} & p_{12} & p_{13} & s_5+\rho\\ p_{21} & p_{22} & s_5-\rho & p_{24}\\ p_{31} & s_5-\rho & p_{33} & p_{34}\\ s_5+\rho & p_{42} & p_{43} & p_{44} \end{vmatrix}=0, \quad (11)$$

which is a fourth degree equation for ρ in which the cubic term is missing:

$$\rho^4+A\rho^2+B\rho+C=0. \quad (12)$$

The coefficients of this equation are not constructed symmetrically however, if they are determined by the solution of the determinant (equation 11). But since ρ is dependent of the mass favored (equation 8), it must be possible to give to the quantities A, B, and C a symmetric form based upon the three masses.

So far no effort has been made to show this *in extenso*. In order to do so, 9 quantities which form a symmetrical system will be selected from the 21 invariables p_{11}, p_{12}, $\cdots p_{66}$, such as p_{11}, p_{33}, p_{55}; p_{22}, p_{44}, p_{66}; p_{12}, p_{34}, p_{56}, or the squares of the relative distances and velocities, and the scalar products of the three vectors of location with their respective velocity vectors. These nine quantities shall be called the *fundamental invariables* of the reduced three-body problem. If these quantities are known as functions of time, all others can be derived from them, some with linear relationships like equations 6 and 7, some

rically, are reflected in the mathematical relationships between the elements of motion. The relative vectors of location of two bodies in reference to the third are $\mathfrak{p}_1$, $\mathfrak{p}_2$, and the relative velocity vectors $\dot{\mathfrak{p}}_1$, $\dot{\mathfrak{p}}_2$; and for convenience let

$$\mathfrak{p}_1=\mathfrak{r}_1, \qquad \dot{\mathfrak{p}}_1=\mathfrak{r}_2, \qquad \mathfrak{p}_2=\mathfrak{r}_3, \qquad \dot{\mathfrak{p}}_2=\mathfrak{r}_4.$$

The geometrical figure formed by these vectors (discounting the aforementioned additional possibilities) is determined by the scalar products possible between them:

$$p_{ab}=(\mathfrak{r}_a\mathfrak{r}_b)=x_ax_b+y_ay_b+z_az_b, \tag{1}$$

which form the matrix

$$\mathfrak{M} = \left\| \begin{matrix} p_{11} & \cdots & p_{14} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ p_{41} & \cdots & p_{44} \end{matrix} \right\|. \tag{2}$$

Because $p_{ab}=p_{ba}$, this matrix will be formed from 10 different elements among which the identity

$$|\mathfrak{M}| = \begin{vmatrix} p_{11} & \cdots & p_{14} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ p_{41} & \cdots & p_{44} \end{vmatrix} = \begin{vmatrix} x_1 & y_1 & z_1 & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_4 & y_4 & z_4 & 0 \end{vmatrix}^2 = 0 \tag{3}$$

holds; therefore only 9 of the quantities (equation 1) which are independent of coordinate transformation are arbitrary, whereas the 10th is a function of the others.

NEW FORM OF THE EQUATIONS OF LAGRANGIAN THEORY

The beauty of the Lagrange theory can be completely expressed only if we provide completely symmetrical formulas. This symmetry cannot be attained if a dominant position is assigned to one of the three bodies, as in the above discussion. Therefore we will introduce a reference system in which none of the vectors is favored. For this purpose, we will call the three masses m_1, m_3, m_5 (fig. 1a) or m_i, m_j, m_k (fig. 1b), where the indexes i, j, k traverse the values 1, 3, 5 or their cyclic permutations. The relative vectors between the masses m_j, m_k are then labeled

$$\mathfrak{r}_i=\overrightarrow{m_jm_k}$$

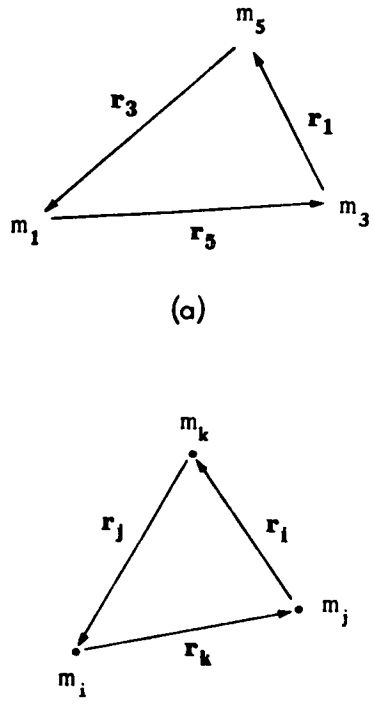


FIGURE 1.—Relative positions of three bodies.

and form the closed train

$$\mathfrak{r}_i+\mathfrak{r}_j+\mathfrak{r}_k=0. \tag{4}$$

The same also applies to the relative velocity vectors; if we set

$$\dot{\mathfrak{r}}_1=\mathfrak{r}_2, \qquad \dot{\mathfrak{r}}_3=\mathfrak{r}_4, \qquad \dot{\mathfrak{r}}_5=\mathfrak{r}_6,$$

or

$$\dot{\mathfrak{r}}_i=\mathfrak{r}_\alpha, \qquad \dot{\mathfrak{r}}_j=\mathfrak{r}_\beta, \qquad \dot{\mathfrak{r}}_k=\mathfrak{r}_\gamma,$$

where $\alpha=i+1$, $\beta=j+1$, $\gamma=k+1$ traverse the values 2, 4, 6 or their cyclic permutations, then it follows from the differentiation of equation 4 that

$$\mathfrak{r}_\alpha+\mathfrak{r}_\beta+\mathfrak{r}_\gamma=0. \tag{5}$$

From the six vectors $\mathfrak{r}_1, \cdots \mathfrak{r}_6$, 36 invariables $(\mathfrak{r}_a\mathfrak{r}_b)=p_{ab}$ can be formed, but only 21 of these are different since $p_{ab}=p_{ba}$. Twelve relationships exist among these quantities—equation 3 and eleven others—by means of which the five quantities p_{15}, p_{25}, $\cdots p_{55}$ and the six quantities p_{16}, p_{66}, $\cdots p_{66}$ can be reduced to the ten invariables of the matrix (equation 2) using the linear rela-

with the help of the Lagrange quantity ρ, which, because of equation 12, will be a function of the fundamental invariables. Twelve of the 21 invariables can therefore be expressed with ρ and the 9 quantities. In actuality, we find

$$\left.\begin{aligned} 2p_{13} &= p_{55}-p_{11}-p_{33}, \\ 2p_{24} &= p_{66}-p_{22}-p_{44}, \\ 2p_{35} &= p_{11}-p_{33}-p_{55}, \\ 2p_{46} &= p_{22}-p_{44}-p_{66}, \\ 2p_{51} &= p_{33}-p_{55}-p_{11}, \\ 2p_{62} &= p_{44}-p_{66}-p_{22}. \end{aligned}\right\} \qquad (13)$$

Further,

$$\left.\begin{aligned} 2s_1 &= p_{12}-p_{34}-p_{56}, \\ 2s_3 &= p_{34}-p_{56}-p_{12}, \\ 2s_5 &= p_{56}-p_{12}-p_{34}; \end{aligned}\right\} \qquad (14)$$

and the remaining 6 invariables are obtained from equation 10.

With the help of these relationships, it is also possible to express the coefficients of the biquadratic equation (equation 12) with the nine fundamental quantities in symmetrical form. We shall present only the results of this somewhat complicated computation. If, for purposes of brevity, we introduce

$$\left.\begin{aligned} \psi_{11} &= p_{11}p_{22}-p_{12}^2, \\ \psi_{13} &= p_{11}p_{44}-p_{34}p_{12}, \\ \psi_{15} &= p_{11}p_{66}-p_{56}p_{12}, \\ \psi_{31} &= p_{33}p_{22}-p_{12}p_{34}, \\ \psi_{33} &= p_{33}p_{44}-p_{34}^2, \\ \psi_{35} &= p_{33}p_{66}-p_{56}p_{34}, \\ \psi_{51} &= p_{55}p_{22}-p_{12}p_{56}, \\ \psi_{53} &= p_{55}p_{44}-p_{34}p_{56}, \\ \psi_{55} &= p_{55}p_{66}-p_{56}^2, \\ X_{13} &= \psi_{13}+\psi_{31}, \\ X_{35} &= \psi_{35}+\psi_{53}, \\ X_{51} &= \psi_{51}+\psi_{15}, \end{aligned}\right\} \qquad (15)$$

we obtain

$$2A = \psi_{11}+\psi_{33}+\psi_{55}-(X_{13}+X_{35}+X_{51}),$$

$$B = \begin{vmatrix} p_{11} & p_{33} & p_{55} \\ p_{22} & p_{44} & p_{66} \\ p_{12} & p_{34} & p_{56} \end{vmatrix} \qquad (16)$$

and

$$4C = A^2-(X_{13}X_{35}+X_{35}X_{51}+X_{51}X_{13})+2(\psi_{11}X_{35}+\psi_{33}X_{51}+\psi_{55}X_{13}).$$

The unhandy necessity of solving equation 12 can be avoided in all practical applications by using the always available integral of the constant impulse moment. Actually, the two known integrals of the reduced three-body problem can be expressed through the fundamental invariables and ρ. We have the energy integral

$$h = \frac{1}{2}\left(\frac{p_{22}}{m_1}+\frac{p_{44}}{m_3}+\frac{p_{66}}{m_5}\right)-\left(\frac{1}{m_1\sqrt{p_{11}}}+\frac{1}{m_3\sqrt{p_{33}}}+\frac{1}{m_5\sqrt{p_{55}}}\right) = \text{constant}; \qquad (17)$$

and, if

$$\psi_{ii} = p_{ii}p_{\alpha\alpha}-p_{i\alpha}^2 \qquad (i = 1, 3, 5;\ \alpha = 2, 4, 6)$$

are the diagonal terms of the matrix ψ_{ik} (equation 15) and if

$$2\phi_{jk} = A + X_{jk} + 2\rho^2,$$

the impulse moment is

$$g^2 = \sum_i \frac{\psi_{ii}}{m_i^2} + 2 \sum_{j,k} \frac{\phi_{jk}}{m_j m_k}$$

or

$$g^2 = G + \frac{m_1 + m_3 + m_5}{m_1 m_3 m_5} (A + 2\rho^2) = \text{constant} . \tag{18}$$

In equation 18 g is the length of the impulse moment vector which stands vertically upon the invariable plane, and

$$G = \frac{\psi_{11}}{m_1^2} + \frac{\psi_{33}}{m_3^2} + \frac{\psi_{55}}{m_5^2} + \frac{X_{35}}{m_3 m_5} + \frac{X_{51}}{m_5 m_1} + \frac{X_{13}}{m_1 m_3} .$$

If the unit of mass is chosen so that $m_1+m_3+m_5=1$, it follows from equation 18 that

$$2\rho^2 = m_1 m_3 m_5 (g^2 - G) - A \tag{19}$$

or, if ρ^2 and ρ^4 are eliminated with the help of equation 19 from equation 12,

$$B\rho = \frac{A^2}{4} - C - \frac{1}{4} m_1^2 m_3^2 m_5^2 (g^2 - G)^2 . \tag{20}$$

Since g^2 is known from the initial conditions of each special problem, ρ can be obtained from equation 19 or, if B is not equal to zero, from equation 20.

The differential equations of the relative three-body problem can be written in vectorial form:

$$\ddot{\mathbf{r}}_i = -\mu_i \mathbf{r}_i + m_i \mathbf{R} , \qquad (i = 1, 3, 5; \ \Sigma m_i = 1) \tag{21}$$

where

$$\left.\begin{aligned} \mu_i &= \frac{1}{r_i^3} , \\ \text{and} \\ \mathbf{R} &= \mathbf{r}_1 \mu_1 + \mathbf{r}_3 \mu_3 + \mathbf{r}_5 \mu_5 \end{aligned}\right\} \tag{22}$$

These differential equations are not independent of one another because of the relationship shown in equation 4. Thus it is sufficient to solve two—perhaps those for r_1, r_3 (two vectorial differential equations of the 2nd order which form a 12th order system). But this system is not symmetrical since one of the three masses, perhaps m_5, is given a favored position. Instead of this, however, the Lagrange theory makes it possible to relate the problem's solution to the integration of nine first order differential equations for the nine fundamental invariables; i..e, a completely symmetrical system with no unnecessary components. In actuality, if we differentiate the quantities

$$p_{ii} = (\mathbf{r}_i \mathbf{r}_i) , \quad p_{i\alpha} = (\mathbf{r}_i \mathbf{r}_\alpha) , \quad p_{\alpha\alpha} = (\mathbf{r}_\alpha \mathbf{r}_\alpha) \qquad (i = 1, 3, 5; \ \alpha = i + 1) \tag{23}$$

with reference to time and eliminate $\dot{\mathbf{r}}_\alpha = \ddot{\mathbf{r}}_i$ by means of equation 21, then

$$\left.\begin{aligned} \frac{1}{2} \dot{p}_{ii} &= p_{i\alpha} , \\ \dot{p}_{i\alpha} &= p_{\alpha\alpha} - p_{ii}\mu_i + m_i p_i , \\ \frac{1}{2} \dot{p}_{\alpha\alpha} &= -p_{i\alpha}\mu_i + m_i p_\alpha , \end{aligned}\right\} \tag{24}$$

in which we abbreviate

$$\left.\begin{aligned} p_i &= (\mathbf{r}_i\mathbf{R}) = p_{i1}\mu_1 + p_{i3}\mu_3 + p_{i5}\mu_5 , \\ p_\alpha &= (\mathbf{r}_\alpha\mathbf{R}) = p_{\alpha 1}\mu_1 + p_{\alpha 3}\mu_3 + p_{\alpha 5}\mu_5 . \end{aligned}\right\} \quad \begin{matrix} (i = 1, 3, 5; \\ \alpha = 2, 4, 6) . \end{matrix} \tag{25}$$

The formation of the quantities p_i and p_α from the fundamental invariables becomes even clearer if we write

$$\left.\begin{aligned} p_i &= p_{ij}(\mu_j - \mu_i) + p_{ik}(\mu_k - \mu_i) , \\ p_\alpha &= s_k(\mu_j - \mu_i) + s_j(\mu_k - \mu_i) + \rho(\mu_k - \mu_j) , \\ 2p_{ij} &= p_{kk} - p_{ii} - p_{jj}; \quad 2s_k = p_{k\gamma} - p_{i\alpha} - p_{j\beta} . \end{aligned}\right\} \tag{26}$$

If equations 24 are integrated (by using, for example, the numerical methods for the given initial conditions), the problem is solved in principle and the integrals (equations 17 and 18) are available to check the results. The Lagrange's differential equation for ρ,

$$\begin{aligned} 2\dot{\rho} &= m_1 p_{35}(\mu_3 - \mu_5) + m_3 p_{51}(\mu_5 - \mu_1) + m_5 p_{13}(\mu_1 - \mu_3) \\ &= \sum m_i p_{jk}(\mu_j - \mu_k) , \end{aligned} \tag{27}$$

can be added to the system (equation 24) as a 10th equation, and is remarkable in its simplicity and symmetry, thereby making possible another thorough check.

Under certain conditions—especially if one of the three bodies has a considerably larger mass than the other two and their orbits can be considered to be disturbed Kepler movements in conical sections—it is useful to introduce, in place of the fundamental invariables, derived invariables which are constructed to correspond with those used in the theory of the two-body problem.* If we set

$$\left.\begin{aligned} \mu_i &= \frac{1}{r_i^3} = p_{ii}^{-3/2} , \\ \sigma_i &= \frac{\dot{r}_i}{r_i} = \frac{p_{i\alpha}}{p_{ii}} , \\ \omega_i &= \frac{p_{\alpha\alpha}}{p_{ii}} , \\ \rho_i &= 2\mu_i - \omega_i , \\ \epsilon_i &= \omega_i - \mu_i , \\ \vartheta_i &= \omega_i - \sigma_i^2 , \end{aligned}\right\} \tag{28}$$

*Stumpff, K., "Calculation of Ephemerides From Initial Values," NASA Technical Note D-1415, in publication 1962.

then these quantities, as derived from equation 24, satisfy the differential equations

$$\left.\begin{aligned}
\dot{\mu}_i &= -3\mu_i\sigma_i \quad \left(\text{or } \dot{r}_i = r_i\sigma_i\right), \\
\dot{\sigma}_i &= \epsilon_i - 2\sigma_i^2 + m_i \frac{p_i}{p_{ii}} , \\
\dot{\omega}_i &= -2\sigma_i\left(\mu_i + \omega_i\right) + 2m_i \frac{p_\alpha}{p_{ii}} , \\
\dot{\rho}_i &= -2\rho_i\sigma_i - 2m_i \frac{p_\alpha}{p_{ii}} , \\
\dot{\epsilon}_i &= -\sigma_i\left(\mu_i + 2\epsilon_i\right) + 2m_i \frac{p_\alpha}{p_{ii}} , \\
\vartheta_i &= -4\eta_i\sigma_i + 2\frac{m_i}{p_{ii}}\left(p_\alpha - p_i\sigma_i\right)
\end{aligned}\right\} \quad \begin{matrix}(i = 1, 3, 5;\\ \alpha = i + 1)\end{matrix} \qquad (29)$$

Since only three of the six quantities (equation 28) are independent, only three of the equation systems (equation 29) need to be considered, i.e., μ_i, σ_i, ϵ_i. If we also set

$$\epsilon_i^* = \epsilon_i + m_1 \frac{p_i}{p_{ii}} , \qquad (30)$$

we obtain the system:

$$\left.\begin{aligned}
\mu_i &= -3\mu_i\sigma_i , \quad \left(\text{or } \dot{r}_i = r_i\sigma_i\right) ; \\
\dot{\sigma}_i &= \epsilon_i^* - 2\sigma_i^2 ; \\
\dot{\epsilon}_i^* &= -\sigma_i\left(\mu_i + 2\epsilon_i^*\right) + \gamma_i ,
\end{aligned}\right\} \qquad (31)$$

where

$$\gamma_i = \frac{m_i}{p_{ii}}\left(2p_\alpha + \dot{p}_i\right) . \qquad (32)$$

The γ_i are three functions of the invariables which, if m_i is one of the small masses, are called perturbations. To show this, we form the equations

$$\begin{aligned}
\dot{r}_i &= r_i\sigma_i , \\
\ddot{r}_i &= \dot{r}_i\sigma_i + r_i\dot{\sigma}_i = r_i\left(\epsilon_i^* - \sigma_i^2\right) , \\
\dddot{r}_i &= \dot{r}_i\left(\epsilon_i^* - \sigma_i^2\right) + r_i\left(\dot{\epsilon}_i^* - 2\sigma_i\dot{\sigma}_i\right) \\
&= r_i\sigma_i\left[3\left(\epsilon_i^* - \sigma_i^2\right) + \mu_i\right] + r_i\gamma_i
\end{aligned}$$

by differentiating the identity $r_i^2=x_i^2+y_i^2+z_i^2$. After elimination of σ_i and $\epsilon_i^*-\sigma_i^2$ from these equations, there are three third order differential equations:

$$r_i\dddot{r}_i + 3\dot{r}_i\ddot{r}_i + \frac{\dot{r}_i}{r_i^2} = r_i^2\gamma_i = m_i(2p_\alpha + \dot{p}_i) , \tag{33}$$

whose nine integrals also solve the reduced problem. Because $\sum p_\alpha=\sum\dot{p}_i=0$, it follows from equation 33 that

$$\sum \frac{1}{m_i}\left(r_i\dddot{r}_i + 3\dot{r}_i\ddot{r}_i + \frac{\dot{r}_i}{r_i^2}\right) = 0 ,$$

an equation whose integral

$$\sum \frac{1}{m_i}\left(r_i\ddot{r}_i + \dot{r}_i^2 - \frac{1}{r_i}\right) = \text{constant}$$

yields the energy law.

CONCLUSION

In the case of general three-body motion where the ratios of mass can have arbitrary values, we always have to rely on numerical integration methods. We have a choice of integrating either the nine first order differential equations (equations 24 or 31) or the three third order differential equations (equations 33). The preferable method is determined by experience and available facilities. The system (equation 33) would probably be preferred for program-controlled electronic computers, since it requires only three, rather than nine, tables of differences.

For instance if the problem is to calculate disturbed planetary orbits, when m_1 denotes the mass of the sun, and m_3, m_5 two small-planet masses (such as Jupiter, $m_3<10^{-3}$ and Saturn, $3m_5<10^{-4}$), and if the perturbations of the second order can be ignored, we use the six equations (equations 24 or 31) or the two equations (equation 33) with $i=3$, 5. The disturbing functions γ_3, γ_5, which are compounded with small factors, are then calculated with the help of the system's undisturbed movement which is known from the initial conditions. This process cannot be used in equations where $i=1$; the motion of the masses m_3 and m_5 relative to one another cannot (as can that of m_3 or m_5 relative to m_1) considered in first approximation to be a Kepler ellipse. It is unnecessary, on the other hand, to solve the equation where $i=1$, since its solution is implicitly a coproduct of the integration process leading to the solution of the other equations.

It should also be mentioned that

$$\left.\begin{aligned} \frac{d}{dt}\left(r_i^2\rho_i\right) &= -2m_i p_\alpha \\ \text{and} \\ \frac{d}{dt}\left(r_i^4\vartheta_i\right) &= 2m_i p_{ii}\left(p_\alpha - p_i\sigma_i\right) \end{aligned}\right\} \tag{34}$$

can be derived from equation 29. These equations are integrable for $m_i=0$ and yield the known integrals of the two-body theory:

$r_i^2\rho_i=1/a_i=$ constant (energy theorem),

$r_i^4\vartheta_i=p_i=$ constant (area theorem).

The task of carrying out the integration of the equations of perturbations according to a method which has been earlier described for the solution of the undisturbed problem will be treated in a later report. An attempt of this type has already been made for Hill's lunar problem,* a particularly simple variation of the restricted three-body problem. The expansion of this result to include the general problem of disturbed orbits will be of considerably more interest.

---

*Stumpff, K., "Remarks on Hill's Lunar Theory, I and II," NASA Technical Notes D–1450, D–1451, in publication 1962.

THE LONG-PERIOD MOTION OF THE TROJANS, WITH SPECIAL ATTENTION TO THE THEORY OF THUERING

KARL STUMPFF*

Goddard Space Flight Center

In 1930, B. Thuering published an approximative theory of long-period motion of the Trojans as a restricted three-body problem. With this theory we can approximate the periodic orbits for all amplitudes in such a favorable manner that the remaining deviations are at most of the order of the mass of Jupiter ($<10^{-3}$). Thuering's solution provides a starting point for an exact theory of the plane long-period Trojan orbits according to the method of the variation of constants. Special attention is devoted to the borderline case, in which the periodic orbits around L_4 and L_5 overlap. This boundary orbit and its adjacent orbits run into the infinitesimal Charlier orbits around L_3—more specifically, into those of the hyperbola type.

INTRODUCTION

The first comprehensive attempt at devising a theory of periodic motions around the Lagrange libration centers L_1, L_2, $\cdots$, L_5 in the restricted problem was undertaken by C. V. L. Charlier around the turn of the century. This was a short time before the discovery, in 1906, of the first Trojan (the planet 588 Achilles) lent practical meaning to this problem which formerly had been of interest only to the theoretician. Charlier's theory merely considered infinitesimal orbits around the libration centers; i.e., orbits whose distances from the libration center remain so small that their squares can be neglected. Charlier showed that there are two families of infinitesimal periodic orbits around L_4 and L_5 (which form an equilateral triangle with the two finite masses Sun and Jupiter) in the rotating coordinate system (in which the Sun and Jupiter hold fixed positions on the x-axis); each of these two families of infinitesimal period orbits consists of a group of concentric, coaxial, and similar ellipses. In each family, the rotation time of all members is equal. The short-period orbits of one family have rotation periods which are only a little longer than those of Jupiter ($T_0=11.86$ years) and which converge on this value, if we allow the Jupiter mass to decrease toward zero. Their ellipticity coefficient $b:a$ deviates only slightly from the boundary value 1:2, toward which it tends when m approaches 0. The minor axis is pointed at the sun (with minor deviations which also disappear when m goes to 0); the major axis thus lies roughly along the tangent of the Jupiter orbit at L_4 (or L_5 respectively).

This family of short-period libration orbits has a very simple meaning which was apparently accorded little attention in the literature. Let us consider a planet moving around the sun in an (undisturbed) elliptical orbit whose semimajor axis and rotation period are equal to those of Jupiter ($a=1$, $T=T_0$) and whose eccentricity e is so small that magnitudes of the order of e^2 can be neglected. Thus the orbit of this body in the rotating system is a small ellipse around a fixed point of a circle of unit radius, its ellipticity coefficient is 1:2, its major axis is tangent to the circle of unit radius, and its rotation period is T_0. The family of the short-period libration orbits around L_4 or L_5 thus converges if we allow the disturbing mass of Jupiter to decrease toward zero toward a group of Kepler ellipses with the same

*NAS–NASA Research Associate; Professor Emeritus, Göttingen University.

rotation time, the same perihelion longitude and the minor eccentricity e, which serves as group parameter.

The orbits of the second family—called Trojan orbits—are concentric to and coaxial with the former. If m is the mass of Jupiter measured in units of solar mass, we have the following approximation: $b:a=(3m),^{1/2}T=T_0(4/27m)^{1/2}$ as the common ellipticity coefficient and the common rotation time of the orbit ellipses. For $m=1/1{,}047$, we get $b:a=1:18.7$ $T=148$ years.

Attempts to extend the theory of the Trojan orbits to noninfinitesimal areas around L_4, L_5 were soon undertaken, such as the work by H. C. Plummer, E. W. Brown, and others.[1,2] These investigations, confined to the consideration of the squares of the planetoid coordinates in relation to the center of libration, revealed a deformation of the orbit ellipses as the distance from the center increased: (1) a slight shift of the center on the circle of unit radius, that is, in the direction away from Jupiter; (2) the symmetry line of the orbit hugs the circle of unit radius; (3) the curvature of the orbit in the greatest elongations on the side facing away from Jupiter is stronger than on the side near Jupiter. On the other hand, the rotation period—considering first-order terms—remains unchanged and begins to increase gradually only as the libration amplitudes grows and as second and higher powers of the coordinates are considered.

AN IMPROVED TROJAN ORBIT

The first attempt at obtaining a clearer picture of the Trojan orbit of arbitrary amplitude width was made as early as 1911 by E. W. Brown,[1] who introduced the polar coordinates r, ϕ (where r is the distance from sun, and ϕ the difference of lengths of the planetoid and Jupiter). And he was able to show that $r-1$ and $\dot{\phi}$ are, at most, of the order of $m,^{1/2}$ while $\dot{r}$ and $\ddot{\phi}$ are at most of the order of m, etc. Thus the orbit lies in the vicinity of a circle of unit radius (fig. 1); and if we use α to designate the average libration amplitude (the average distance of the Trojan from the libration center in the maximum elongations), then $\alpha(m)^{1/2}$ is the order of magnitude of the maximum distance of the Trojan from the circle of unit radius. In addition, Brown succeeded in estimating the distances of the Trojan from the libration center in the elongations; in particular he found that, if $\phi_0=60°$ is the length of L_4 (related to the longitude of Jupiter), the longitude of the longest elongation on the side facing away from Jupiter approaches 180°, while that of the elongation near Jupiter decreases to about 24°. We thus have a boundary orbit in the family of the Trojan orbits around L_4 which, in the elongation facing away from Jupiter, reaches to the opposite point of Jupiter relative to the Sun and there meets the corresponding boundary orbit around L_5, although the elongation of the boundary orbits on the side near Jupiter remains separated from Jupiter by a longitude difference of about 24°. The elongations of the boundary orbits of L_4 therefore are 120° and 36°, respectively, and the deviations of these orbits from the circle of unit radius are of the order of $m^{1/2}\approx\frac{1}{10}$.

The motion theory of the real Trojans, of which nine around L_4 and five around L_5 have become known thus far, has been worked out by various authors, of whom we shall mention only E. W. Brown[3] and A. Wilkens.[4–7] These theories concern the spatial movement of these planets, including all perturbations due to Jupiter and other planets, especially Saturn. These orbits are not periodic. In the plane restricted problem—a problem of the fourth order represented by two differential equations of the second order, every libration around L_4 (or L_5) is composed of two independent motions which have the characteristics of the two Charlier families (expanded to noninfinitesimal regions). In this problem we have periodic Trojan

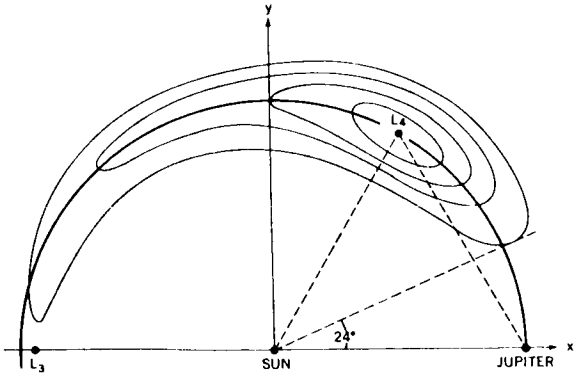


FIGURE 1.—Family of periodic Trojan orbits. The deviation of the curves from the unit circle is exaggerated.

orbits with rotation times (equal to or greater than 148[a]) obtained by setting the initial conditions so that the short-period component vanishes. Thus these periodic Trojan orbits are solutions of a second-order system.

THUERING'S WORK ON LONG-PERIOD LIBRATIONS

Owing to Thuering's work,[8] we now have an approximated theory of these long-period librations. Simply and clearly, this theory gives intermediate orbits, whose deviation from the exact-period orbits does not exceed the order of m even in the boundary case. Thus these orbits can be used as approximations for an exact theory and therefore provide a very clear picture of the shape of the orbits and the form of motion within them.

In a plane coordinate system, x_1, y_1 are the rectangular, heliocentric coordinates of Jupiter which moves around the sun in a circular orbit (radius $a_1=1$). Thus, for Jupiter, we have:

$r_1 = 1$, the radius vector $=5.20$ A.E.,
$\lambda_1 = n_1(t-t_0)$, the mean length,
$l_1 = \lambda_1$, the true length;

and we may let $n_1=(1+m)^{1/2}$ for the mean motion in the time unit, if we set this time unit as equal to $(5.20)^{3/2}/k$ average days. We also let x, y be the rectangular coordinates and r, l the polar coordinates of the Trojan.

As elements of the osculatory orbit of the Trojan for any time t, we introduce:

$a=$major semiaxis,
$e=\sin\phi=$numerical eccentricity,
$\omega=$length of the perihelion.

Furthermore, let M be the mean anomaly, so that $l=\lambda+2e\sin M+$terms of higher order in e, where $\lambda=n(t-t_0)=a^{-3/2}\ (t-t_0)$ is the mean longitude of the planetoid.

Using the canonic elements of Poincaré,

$$L = \sqrt{a}, \qquad \lambda = M+\omega,$$

$$p = 2\sqrt{a}\sin^2(\phi/2), \qquad q = -\omega,$$

we obtain the following differential equations for the motion:

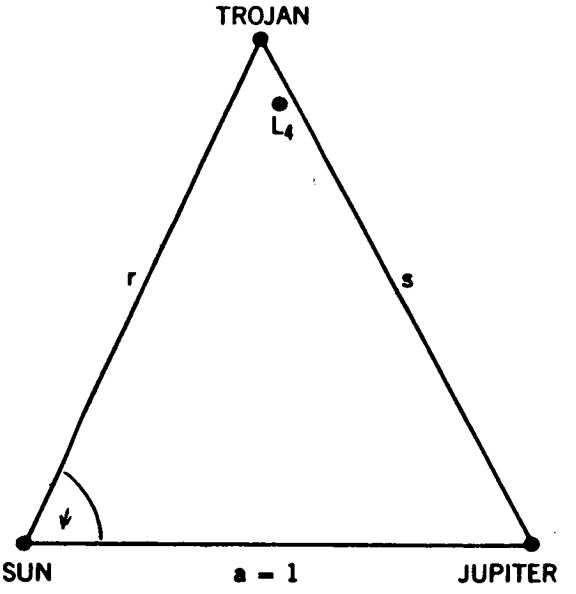


FIGURE 2.—The triangle sun–Jupiter–Trojan.

$$\frac{dL}{dt} = \frac{\partial F}{\partial \lambda}, \qquad \frac{dp}{dt} = \frac{\partial F}{\partial q},$$

$$\frac{d\lambda}{dt} = -\frac{\partial F}{\partial L}, \qquad \frac{dq}{dt} = -\frac{\partial F}{\partial p}, \tag{1}$$

where the Hamilton function

$$F = \frac{1}{2a} + m\left[\frac{1}{s} - (xx_1 + yy_1)\right].$$

From figure 2 we have

$$s^2 = 1+r^2-2r\cos\Psi,$$
$$xx_1+yy_1 = r\cos\Psi,$$
$$\Psi = l-l_1$$
$$= \lambda-\lambda_1+2e\sin M+\cdots$$
$$= \alpha+2e\sin M+\cdots.$$

Now, in the osculatory orbit Brown showed that

$$\dot{r} = \frac{e\sin v}{\sqrt{a(1-e^2)}},$$

where $\dot{r}$ is, at most, of the order of m (for $v=90°$). From this it follows that e is also the order of m. Therefore we shall develop F into a power series of e,

$$F = F_0 + eF_1 + e^2F_2 + \cdots,$$

where F_0 is free of the eccentricity e. However, $F=F(r,\psi)$, for, if we set up $a=a_1(1+\rho)=1+\rho$, $r=a(1-e\cos E)$, $\psi=\alpha+2e\cos M+\cdots$,

$$F = F(\rho, \alpha, e\cos E, e\cos M, \cdots) .$$

And for $e=0$ we have

$$\begin{aligned} F_0 &= F_0(\rho, \lambda) \\ &= F_0(\rho, \alpha) \\ &= \frac{1}{2(1+\rho)} + m\left[\frac{1}{\sqrt{2(1+\rho)(1-\cos\alpha)+\rho^2}} - (1+\rho)\cos\alpha\right] . \end{aligned} \tag{2}$$

The differential equations 1 are limited to the second order system, if we neglect magnitudes of the order of e, and of the order of m:

$$\frac{dL}{dt} = \frac{\partial F_0}{\partial \lambda} , \qquad \frac{d\lambda}{dt} = -\frac{\partial F_0}{\partial L} ,$$

or if we set up

$$\begin{aligned} L &= \sqrt{1+\rho} , \\ \lambda &= \lambda_1 + \alpha \\ &= \alpha + \sqrt{1+m}\,(t-t_0) , \end{aligned}$$

they are limited to

$$\begin{aligned} \frac{d\rho}{dt} &= 2\sqrt{1+\rho}\,\frac{\partial F_0}{\partial \alpha} , \\ -\frac{d\alpha}{dt} &= \sqrt{1+m} + 2\sqrt{1+\rho}\,\frac{\partial F_0}{\partial \rho} . \end{aligned} \tag{3}$$

The remaining two equations of equation 1 are now superfluous, since p vanishes and $q=-\omega$ becomes meaningless when $e=0$.

If in equation 2 we now ignore all terms of the order m^2 (noting that ρ^2 is of the order of m), we get

$$F_0 = \frac{1}{2(1+\rho)} + m\left[\frac{1-\frac{\rho}{2}}{2\sin\frac{\alpha}{2}} - (1+\rho)\cos\alpha\right]$$

and

$$\begin{aligned} \frac{\partial F_0}{\partial \alpha} &= m\left[\frac{\frac{\rho}{2}-1}{4\sin^2\frac{\alpha}{2}}\cos\frac{\alpha}{2} + (1+\rho)\sin\alpha\right] \\ &= m\sin\alpha\left[1 - \frac{1}{8\sin^3\frac{\alpha}{2}} + \rho\left(1 + \frac{1}{16\sin^3\frac{\alpha}{2}}\right)\right], \end{aligned}$$

$$\frac{\partial F_0}{\partial \rho} = -\frac{1}{2(1+\rho)^2} - m\left(\frac{1}{4\sin\frac{\alpha}{2}} + \cos\alpha\right) .$$

From the preceding two equations and equations 3, we get

$$\frac{d\rho}{dt} = 2m\sqrt{1+\rho}\sin\alpha\left[1 - \frac{1}{8\sin^3\frac{\alpha}{2}} + \rho\left(1 + \frac{1}{16\sin^3\frac{\alpha}{2}}\right)\right] \tag{4}$$

and

$$\frac{d\alpha}{dt} = (1+\rho)^{-3/2} - \sqrt{1+m} + 2m\sqrt{1+\rho}\left(\frac{1}{4\sin\frac{\alpha}{2}} + \cos\alpha\right) \tag{5}$$

If we differentiate equation 5 again with respect to time and replace $d\rho/dt$ with equation 4 and then ignore all terms of the order of m^2, we obtain

$$\frac{d^2\alpha}{dt^2} = -3m(1+\rho)^{-2}\sin\alpha\left[1 - \frac{1}{8\sin^3\frac{\alpha}{2}} + \rho\left(1 + \frac{1}{16\sin^3\frac{\alpha}{2}}\right)\right]$$
$$-2m\sqrt{1+\rho}\sin\alpha\left(1 + \frac{1}{16\sin^3\frac{\alpha}{2}}\right)\frac{d\alpha}{dt} . \tag{6}$$

But according to equation 5, we have

$d\alpha/dt = -\frac{3}{2}\rho +$ terms of the order of m, and also, ρ^2, (7)

ρ and $d\alpha/dt$ thus, as we said before, are of the same order m.[1/2] If we now ignore in equation 6 all terms of the order m^2, $m\rho^2$, $m\rho(d\alpha/dt)$, we may, from equation 7, let

$$\rho = -\frac{2}{3}\frac{d\alpha}{dt} \tag{8}$$

and instead of the two differential equations of the first order (equations 4 and 5), we obtain one differential equation of the second order of $\alpha(t)$:

$$\frac{d^2\alpha}{dt^2} + f(\alpha)\left(\frac{d\alpha}{dt} + \frac{3}{4}\right) = 0$$

where

$$f(\alpha) = 4m\sin\alpha\left(1 - \frac{1}{8\sin^3\frac{\alpha}{2}}\right) \tag{9}$$

The solution of equation 9 also determinates ρ from equation 8 (except for terms of the order m).

The integration of equation 9 can be accomplished in the following elegant manner. We have a known function of α,

$$\phi(\alpha) = -\int f(\alpha)\,d\alpha = 4m\left(\cos\alpha - \frac{1}{2\sin\frac{\alpha}{2}}\right).$$

Now, if we let

$$\frac{d\alpha}{dt} = \phi(\alpha) + \gamma(\alpha) , \tag{10}$$

it suffices to determine $\gamma(\alpha)$, for afterwards $\rho(t)$ is obtained from equation 8 and $\alpha(t)$ by squaring equation 10. From equation 10, it follows that

$$\frac{d^2\alpha}{dt^2} = \frac{d\phi}{dt} + \frac{d\gamma}{dt}$$
$$= \left(\frac{d\phi}{d\alpha} + \frac{d\gamma}{d\alpha}\right)\frac{d\alpha}{dt}$$
$$= \left[\frac{d\gamma}{d\alpha} - f(\alpha)\right]\frac{d\alpha}{dt} .$$

If we insert the above equation into equation 9, we get

$$\frac{d\gamma}{d\alpha}\frac{d\alpha}{dt} + \frac{3}{4}f(\alpha) = 0$$

or

$$\frac{d\gamma}{d\alpha} = -\frac{3}{4}\frac{f(\alpha)}{\phi(\alpha) + \gamma(\alpha)} \quad . \tag{11}$$

If, for the moment, we set

$$\phi + \gamma = \frac{3}{4u} \quad , \tag{12}$$

we find that

$$\frac{d}{d\alpha}(\phi + \gamma) = -f(\alpha) + \frac{d\gamma}{d\alpha}$$

$$= -\frac{3}{4u^2}\frac{du}{d\alpha} \; ;$$

or, if from equation 11, we set

$$\frac{d\gamma}{d\alpha} = -uf(\alpha) \; ,$$

we find that

$$\frac{du}{d\alpha} = \frac{4}{3} f u^2 (1 + u) \; ,$$

i.e.,

$$\frac{du}{u^2(1+u)} = \frac{4}{3} f(\alpha)\, d\alpha$$

$$= -\frac{4}{3} d\phi \quad .$$

This equation is integrable and, if ϕ_0 is an arbitrary constant, we obtain

$$\frac{4}{3}(\phi - \phi_0) = \frac{1}{u} + ln\frac{u}{1+u} \quad .$$

If, instead of u, we again introduce γ from equation 12, we get

$$-\frac{4}{3}(\phi_0 + \gamma) = ln\frac{1}{1 + \frac{4}{3}(\phi + \gamma)} \; ,$$

or

$$e^{-(4/3)(\phi_0+\gamma)}\left[1 + \frac{4}{3}(\phi+\gamma)\right] = 1 \; , \tag{13}$$

an equation which determines $\gamma = \gamma(\phi, \phi_0)$. Since from equation 10, ϕ is of the order of m and γ is of the order m,$^{1/2}$ we can solve equation 13 by using a rapidly converging power series. From equation 13 and

$$y = \frac{4}{3}\frac{d\alpha}{dt} = \frac{4}{3}(\phi + \gamma),$$

$$z = \frac{4}{3}(\phi - \phi_0) \; ,$$

we have

$$e^{-y}(1 + y) = e^{-z} \; ,$$

i.e.,

$$1 - \frac{1}{2}y^2 + \frac{1}{3}y^3 - \frac{1}{8}y^4 + \cdots = 1 - z + \frac{1}{2}z^2 - \cdots \; ; \tag{14}$$

and to a first approximation,

$$y^2 \approx 2z = \zeta^2 \; ,$$

where

$$\zeta = \pm\sqrt{2z} \; .$$

Therefore we can set up a power series of ζ for y and by comparing coefficients we obtain the following from equation 14:

$$y = \pm\zeta - \frac{1}{3}\zeta^2 \pm \frac{1}{36}\zeta^3 - \frac{1}{135}\zeta^4 \pm \cdots$$

or, if we neglect terms in ζ^4 and higher terms (which are at least of the order of m^2) we get

$$y = \frac{4}{3}\frac{d\alpha}{dt}$$

$$\approx \pm\zeta\left(1 \mp \frac{1}{6}\zeta\right)^2 \quad .$$

We now have the approximation

$$\frac{d\alpha}{dt} = \pm\frac{3}{4}\zeta\left(1 \mp \frac{1}{6}\zeta\right)^2$$

$$= \pm\sqrt{\frac{3}{2}(\phi - \phi_0)}\left[1 \mp \frac{1}{3}\sqrt{\frac{2}{3}(\phi - \phi_0)}\right]^2 \quad . \tag{15}$$

where for real orbits

$$\phi = 4m\left(\cos\alpha - \frac{1}{2\sin\frac{\alpha}{2}}\right) \geqq \phi_0 \quad . \qquad (16)$$

THE MOTION OF THE TROJANS

The motion takes place in such a manner that α, the longitudinal difference between Trojan and Jupiter, fluctuates between two limits $\alpha_1 \times \alpha_2$. These boundary points, which are return points between the clockwise and counterclockwise motions with respect to L_4, are given by $d\alpha/dt = 0$, (i.e., $\phi = \phi_0$). This equation, when $x = \sin(\alpha/2)$, can also be written in the algebraic form

$$x^3 - \frac{1}{2}x\left(1 - \frac{\phi_0}{4m}\right) + \frac{1}{4} = 0 \quad . \qquad (17)$$

Equation 17 has, at most, two real roots in the range $0 < \alpha \times 180°$, that is: for $0 < x \times 1$; for $\phi_0 = -2m$ the double root is $x = \frac{1}{2}$, $\alpha = 60°$. For this value of the parameter ϕ_0, α is constant, therefore the planetoid remains constantly in L_4. For the boundary orbit, we have the maximum elongation $x = 1$ in the side away from Jupiter when $\alpha_1 = 180°$ (i.e., for $\phi_0 = -6m$). The corresponding second root of Equation 17 results from $\alpha_2 = 23{:}9$. Figure 3 shows the position of the return points for different values of ϕ_0 between $-2m$ and $-6m$.

The maximum angular speed of the Trojan relative to L_4 occurs when $d^2\alpha/dt^2 = 0$, because, from equation 9, $f(\alpha) = 0$, since $d\alpha/dt + \frac{3}{4}$ is not equal to zero. But this condition is met for sin $\alpha/2 = \frac{1}{2}$ (i.e., $\alpha = 60°$). The angular speed thus

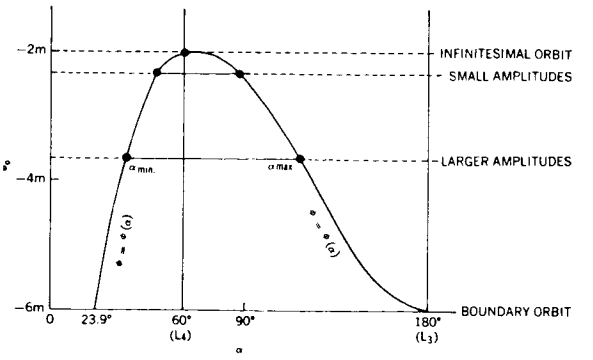


FIGURE 3.—Limits of α, for Trojan orbits of different amplitude.

reaches its maximum if the planetoid, seen from the sun, passes the libration center. When $\alpha = 60°$ we get $\phi = -2m$ from equation 16 and hence, we get $\phi - \phi = 4m$ for case of the boundary orbit ($\phi_0 = -6m$). If we insert this result in equation 15 and if we consider that the time unit was so selected that $T_0 = 2\pi$, and if we use the mean day as the unit of time, we get for the maximum angular speed in the boundary orbit $+23''.4$ and $-21''.9$ daily, respectively.

Since ρ is proportional to $-\dot{\alpha}$, it follows that the motion occurs in the positive sense (counterclockwise) when the Trojan traverses that part of its orbit which lies within the orbit of Jupiter. The motion around L_4 (this also applies to L_5) is thus counterclockwise, and the angular speed is somewhat greater on the inside than on the outside part. Thuering gives a rougher approximation to the solution of equation 14 and obtains $= \dot{\alpha}_{max}$ $\pm 22''.6$, as a medium value.

The rotation time in the periodic orbit can be obtained through integration of equation 15. From

$$dt = \pm\frac{4}{3}\frac{d\alpha}{\zeta\left(1 \pm \frac{1}{6}\zeta\right)^2}, \quad \zeta = \sqrt{\frac{8}{3}(\phi - \phi_0)}, \quad \phi = \phi(\alpha)$$

we get

$$t - t_0 = \pm\frac{4}{3}\int_{\alpha_0}^{\alpha}\frac{d\alpha}{\zeta\left(1 \pm \frac{1}{6}\zeta\right)^2} \quad ,$$

which means that the time for the inside orbit arc is

$$P_1 = \frac{4}{3}\int_{\alpha_{min}}^{\alpha_{max}}\frac{d\alpha}{\zeta\left(1 + \frac{1}{6}\zeta\right)^2} \quad ;$$

and the time for the outside orbit arc is

$$P_2 = \frac{4}{3}\int_{\alpha_{min}}^{\alpha_{max}} \frac{d\alpha}{\zeta\left(1-\frac{1}{6}\zeta\right)^2} .$$

Therefore the total rotation time is $P=P_1+P_2$.

Since the motion of the Trojan is always very slow even in the boundary orbit, we find, on the basis of the Jacobi integral $y^2=2\Omega-C$, that the orbits are always near the Hill zero speed curves: $2\Omega-C=0$ and approach these curves particularly closely at the return points. Every Trojan orbit is therefore enclosed between two neighboring Hill curves with slightly differing values of C. From this we can conclude that the boundary orbit, whose outermost limit is at $\alpha=180°$, will run into that Hill curve which represents the boundary case between the isolated, bean-shaped curves around L_4 and L_5 and the horse-shoe-shaped curves which surround the two libration centers. This boundary curve has a double point in L_3, the libration center which is situated near the point of the x-axis ($x=-1$, $y=0$) opposite Jupiter, and has the abscissa $x=-1+(\frac{7}{12})m$ (except for the terms of the order of m^2).

CONCLUDING REMARKS

According to the Charlier theory we have two families of orbits around L_3, as well as around all other libration centers; one of these families is periodic and, in the infinitesimal range, consists of concentric, coaxial, and similar ellipses whose ellipticity ratio for $m\leftarrow 0$ is transformed into $b:a=1:2$ and the rotation time becomes T_0. Like the family of the short-period orbits around L_4 and L_5, this family represents slightly eccentric Kepler ellipses which are little disturbed by Jupiter. The second family consists of nonperiodic orbits, which can be represented in the vicinity of L_3 by a group of concentric, coaxial, and similar hyperbolas or, rather, by two groups whose common asymptotes touch the above mentioned Hill boundary curves in L_3. These conditions are shown in figure 4; there is no doubt that the pointed hyperbolas in the upper and lower sector of the figure belong to those Trojan orbits around L_4 and L_5, which in the boundary orbit approach each other to infinitesimal intervals. The boundary orbit itself makes the junction at L_3 and is represented by the asymptote pair in the infinitesimal range. Thuering's theory, which provides only an approximation, also gives $\rho=0$, for $\dot{\alpha}=0$ (i.e., the boundary orbit according to this theory does not make the junction at L_3, but at the opposite point $(-1, 0)$ of Jupiter). The deviation however is of the order m—the order of the terms neglected in Equation 7.

The second group of hyperbolas, whose shape is flat, and which fills the right and left sector of figure 4, has as common main axis the x-axis itself, and all these hyperbolas thus intersect the x-axis at right angles. They are tangent there to those periodic orbits which jointly enclose L_4 and L_5 and must be considered as continuations of the two groups of Trojan orbits beyond their common boundary orbit. We can use the term "Transtrojans" to describe the planetoids which

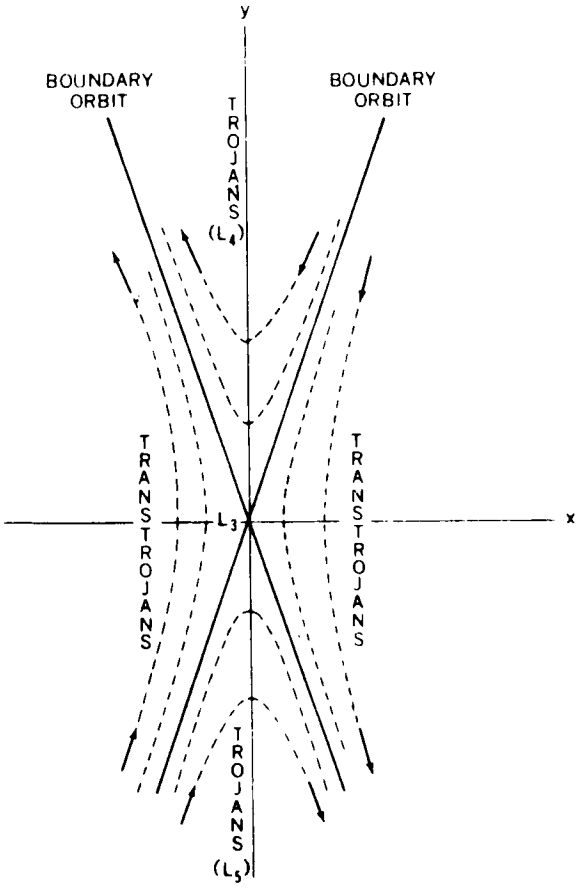


FIGURE 4.—Infinitesimal region around L showing the Trojans, Boundary orbit and Transtrojans.

would run on such horse-shoe-shaped orbits. As a matter of fact, these Transtrojan orbits are among the periodic orbits of the restricted problem—and we can obtain them by numerical integration of this problem's differential equations, if we start from a point on the x-axis which is near L_3 at slow speed perpendicular to the x-axis; and if we vary the speed until the orbit turns back on itself.

Thuering's later attempt to arrive at a periodic orbit starting from the point $(-1, 0)$ when $\dot{y}=0$, therefore was unsound and necessarily unsuccessful.

REFERENCES

1. BROWN, E. W., "On a New Family of Periodic Orbits in the Problem of Three Bodies," *Monthly Not. Roy. Astronom. Soc.*, **71**(5): 438–454, Mar. 10, 1911.
2. ———, "On the Oscillating Orbits about the Triangular Equilibrium Points in the Problem of Three Bodies," *Monthly Not. Roy. Astronom. Soc.*, **71**(6): 492–502, Apr. 12, 1911.
3. ———, "The General Orbits of the Asteroids of the Trojan Group," *Astronom. J.*, **35**(9–10): 69–80, 124, August 1923.
4. WILKENS, A., "Methoden zur Ermittelung der speziellen und der absoluten Koordinatenstorungen der Planeten der Jupitergruppe durch Jupiter," *Astronomische Nachrichten*, **205**(4906): 145–156, Oct. 18, 1917.
5. ———, "Uber die Integration der Differentialgleichungen der Koordinatenstorungen der Planeten der Jupitergruppe," *Astronomische Nachrichten*, **206**(4937): 149–156, Apr. 20, 1918.
6. ———, "Zur Theorie der Librationen im Sonnensystem," *Astronomische Nachrichten*, **227**(5445): 337–348, May 15, 1926.
7. ———, "Untersuchungen zur Theorie der Jupitergruppe," *Astronomische Nachrichten*, **245**(5870–5971): 229–264, May 4, 1932.
8. THURING, B., "Uber die Planeten vom Kommensurabilitass typus 1:1," *Astronomische Nachrichten*, **238**: 357–368, 1930.
9. RABE, E., "Determination and Survey of Periodic Trojan Orbits in the Restricted Problem of Three Bodies," *Astronom. J.*, **66**(9): 500–513, November 1961.

REMARKS ON HILL'S LUNAR THEORY, PART I

KARL STUMPFF

Goddard Space Flight Center

G. W. Hill, by neglecting solar parallax and the eccentricity of the solar orbit, presented a particularly simple form of the differential equation of the restricted three-body problem for the motion of a massless satellite. In the present paper, the Hill equation is modified to give a third order differential equation for r—the planetocentric distance of the satellite. And this equation can be solved by iteration (if the satellite orbit is considered to be a disturbed Kepler ellipse). This new solution is not only suitable for computer applications but it is independent of the coordinate system and is valid for both a fixed and a rotating system whereas Hill's solution was limited to simple-periodic orbits in a rotating system. The discussion of Hill's lunar problem provides a simple example of a method applicable also to more difficult problems such as the restricted three-body problem, in which conic-section orbits can be considered good approximate solutions.

INTRODUCTION

By discounting the solar parallax, G. W. Hill[1] gave a particularly simple form to the differential equation of the restricted three-body problem for the motion of a massless satellite around a planet in a circular orbit about the sun. This relatively simple problem offers tempting considerations and experiments, which will be discussed here. In this paper, it will be shown that the Hill equation can be modified so that a differential equation of the third order for r—the planetocentric distance of the satellite—results, in which (besides r and its derivations through the third order) only the Jacobian constant occurs. And this equation can be solved by iteration if the satellite orbit is considered to be a disturbed Kepler ellipse.

HILL'S DIFFERENTIAL EQUATIONS

Hill's differential equations can be written as:

$$\left.\begin{aligned} \ddot{\xi} - 2\dot{\eta} &= -\xi\left(\frac{1}{r^3} - 3\right), \\ \ddot{\eta} + 2\dot{\xi} &= -\frac{\eta}{r^3}, \end{aligned}\right\} \quad (1)$$

where

$$r^2 = \xi^2 + \eta^2 .$$

Here ξ, η are the right-angle planetocentric coordinates of the assumed massless satellite S in a coordinate system (fig. 1) which revolves with the constant angular velocity $n=1$. And we have

$$n = k\sqrt{\frac{m_1 + m}{a^3}} = 1 \quad (2)$$

for the mean motion of the planet P in its circular orbit around the sun S, which is to be assumed at a great distance, a, along the negative ξ-axis.

The units of mass are to be so selected that equation 2 is valid. This is the case when we set $k=1$ and $a^3=m_1+m$ where m_1 is the mass of the sun and m the mass of the planet. In addition, the

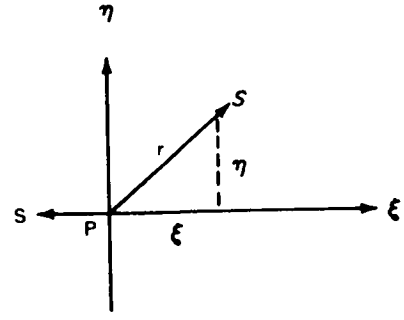


FIGURE 1.—Geometric representation of a massless satellite S of a planet P.

unit of mass is selected so that $m=1$. In the case of the motion of a satellite, or moon, about the earth, $m_1 \approx 330{,}000$ and the unit of time is equal to $1/k = 58.^{d}13244$. It also follows that $a = (m_1+m)^{1/3} \approx 69$. Thus equation 1 is obtained if, in the differential equation of the restricted three-body problem, m_1 and a are allowed to go towards infinity so far that $m_1/a^3 \to 1$, but m_1/a^5, m_1/a^7, $\cdots \to 0$. This simplification of the problem means that the sun, as a disturbing body, is removed to any great distance desired (so that its parallax at the earth-moon distance becomes unnoticeably small), but that its disturbing effect upon the satellite's motion (which is proportional $m_1 a^{-3}$) about the earth remains unchanged.

Hill's equations (equation 1) cannot simply be transformed into the equations of the undisturbed motion of the satellite, since the mass of the disturbing body (the sun) does not occur explicitly in them, but rather is implicitly contained in the definition of the unit of length. The equations can, however, be given a form which makes it possible to combine the disturbed and undisturbed motion. If we write

$$\left.\begin{aligned} \ddot{\xi} - 2\dot{\eta} &= \xi\left(1 - \frac{1}{r^3} + 2k\right), \\ \ddot{\eta} + 2\dot{\xi} &= \eta\left(1 - \frac{1}{r^3} - k\right), \\ \xi^2 + \eta^2 &= r^2 \end{aligned}\right\} \qquad (3)$$

where

then these equations assume the form of equation 1 for $k=1$, whereas letting $k=0$ yields the equations of undisturbed motion (in a coordinate system revolving with the angular velocity $n=1$).

Equations 3 have the Jacobian integral

$$\dot{\xi}^2 + \dot{\eta}^2 = r^2(1-k) + \frac{2}{r} + 3k\xi^2 - C . \qquad (4)$$

THE DIFFERENTIAL EQUATION FOR *r(t)*

The two differential equations of the second order (equation 3) form a fourth-order system. In place of them, a single fourth-order differential equation can be written for a suitable variable—perhaps $r = (\xi^2+\eta)^{21/2}$. With the help of the Jacobian integral, the order of this equation can be decreased by one, so that we have a relation of the form

$$f(\dddot{r}, \ddot{r}, \dot{r}, r; C) = 0 . \qquad (5)$$

This same consideration applies *mutatis mutandis* for the complicated *problème restreint*, which we will discuss, but first we will make this experiment with the simple Hill problem.

If we differentiate

$$r^2 = \xi^2 + \eta^2 \qquad (6)$$

three times, and if we eliminate $\ddot{\xi}$, $\ddot{\eta}$ by using equation 3 and $\dot{\xi}^2+\dot{\eta}^2$ by using equation 4, we obtain

$$\left.\begin{aligned} r\dot{r} &= \xi\dot{\xi} + \eta\dot{\eta} , \\ r\ddot{r} + \dot{r}^2 &= 2(\xi\dot{\eta} - \eta\dot{\xi}) + 2r^2(1-k) + \frac{1}{r} + 6k\xi^2 - C , \\ r\dddot{r} + 3\dot{r}\ddot{r} &= 6k(2\xi\dot{\xi} - \xi\eta) - \frac{\dot{r}}{r^2} - 4kr\dot{r} . \end{aligned}\right\} \tag{7}$$

Equations 6, 7 and 4 form a system

$$\phi_0 = r^2 = \xi^2 + \eta^2 , \tag{8a}$$

$$\phi_1 = r\dot{r} = \xi\dot{\xi} + \eta\dot{\eta} , \tag{8b}$$

$$\phi_2 = r\ddot{r} + \dot{r}^2 - \frac{1}{r} - 2r^2(1-k) + C = 2(\xi\dot{\eta} - \eta\dot{\xi}) + 6k\xi^2 , \tag{8c}$$

$$\phi_3 = r\dddot{r} + 3\dot{r}\ddot{r} + \frac{\dot{r}}{r^2} + 4kr\dot{r} = 6k(2\xi\dot{\xi} - \xi\eta) , \tag{8d}$$

$$\phi_4 = r^2(1-k) + \frac{2}{r} - C = \dot{\xi}^2 + \dot{\eta}^2 - 3k\xi^2 , \tag{8e}$$

in which five functions $\phi_0, \cdots \phi_4$ of r, $\dot{r}$, $\ddot{r}$, $\dddot{r}$, and C appear on the left and appear on the right as functions of the four quantities ξ, η, $\dot{\xi}$, $\dot{\eta}$. Therefore, to obtain the desired equation in the form of equation 5, it suffices to eliminate the latter four quantities from equation 8.

ELIMINATION

It follows from equations 8b and 8c, and by taking equation 8a, that

$$2\phi_0\dot{\xi} = 2\phi_1\xi - \phi_2\eta + 6k\xi^2\eta \tag{9a}$$

$$2\phi_0\dot{\eta} = 2\phi_1\eta + \phi_2\xi - 6k\xi^3 . \tag{9b}$$

If we multiply equation 9a by $\dot{\xi}$, equation 9b by $\dot{\eta}$, and add, we get

$$2\phi_0(\dot{\xi}^2 + \dot{\eta}^2) = 2\phi_1^2 + (\xi\dot{\eta} - \eta\dot{\xi})(\phi_2 - 6k\xi^2)$$

or, since from equation 8c it follows that $\xi\dot{\eta}-\eta\dot{\xi}=\frac{1}{2}(\phi_2-6k\xi^2)$,

$$\phi_0(\dot{\xi}^2 + \dot{\eta}^2) = \phi_1^2 + \frac{1}{4}(\phi_2 - 6k\xi^2)^2 . \tag{10}$$

If equations 9 and 10 are substituted into equations 8d and 8e, there results

$$\phi_0\phi_3 = 6k\left[2\phi_1\xi^2 - \xi\eta(\phi_0 + \phi_2 - 6k\xi^2)\right] , \tag{11a}$$

$$\phi_0\phi_4 = \phi_1^2 + \frac{1}{4}(\phi_2 - 6k\xi^2)^2 - 3k\phi_0\xi^2 , \tag{11b}$$

and from equation 8a

$$\phi_0 \xi^2 = \xi^4 + (\xi\eta)^2 . \tag{11c}$$

The $6k\xi^2$ and the $6k\xi\eta$ terms can easily be eliminated from equations 11 by letting $x=6k\xi^2$, $y=6k\xi\eta$, for if we set

$$\phi_0 + \phi_2 - x = \lambda , \tag{12}$$

from equation 11b we get

$$4(\phi_0\phi_4 - \phi_1^2) = (\lambda - \phi_0)^2 - 2\phi_0(\phi_0 + \phi_2 - \lambda) = \lambda^2 + \phi_0^2 - 2\phi_0(\phi_0 + \phi_2) .$$

The above equation means that

$$\lambda^2 = \phi_0^2 + 2\phi_0(\phi_2 + 2\phi_4) - 4\phi_1^2 \tag{13}$$

is dependent only upon ϕ_0, ϕ_2, ϕ_4; i.e., upon r, $\dot{r}$, $\ddot{r}$, C.

From equations 11a, 11c, and 12 we have

$$y\lambda = 2\phi_1(\phi_0 + \phi_2 - \lambda) - \phi_0\phi_3 ,$$

$$y^2 = -(\phi_0 + \phi_2 - \lambda)^2 + 6k\phi_0(\phi_0 + \phi_2 - \lambda) ,$$

from which, we obtain a fourth-order equation in λ by eliminating y:

$$\lambda^4 + a_1\lambda^3 + a_2\lambda^2 + a_3\lambda + a_4 = 0 \tag{14}$$

where the coefficients of μ are dependent only upon $\phi_0, \cdots \phi_4$:

$$\left.\begin{aligned} a_1 &= 2\left[\phi_0(3k-1) - \phi_2\right] , \\ a_2 &= 4\phi_1^2 - 9k^2\phi_0^2 + \tfrac{1}{4}a_1^2 , \\ a_3 &= 4\phi_1\left[\phi_0\phi_3 - 2\phi_1(\phi_0 + \phi_2)\right] , \\ a_4 &= \left[\phi_0\phi_3 - 2\phi_1(\phi_0 + \phi_2)\right]^2 . \end{aligned}\right\} \tag{15}$$

It is worthy of note that $\ddot{r}$ occurs only in ϕ_3, specifically in the combination $\phi_0\phi_3 - 2\phi_1(\phi_0+\phi_2)$, which, in turn, appears only in α_3 and α_4. Therefore, if we set

$$f = \phi_0\phi_3 - 2\phi_1(\phi_0 + \phi_2) , \tag{16}$$

we can also write equation 14 as

$$f^2 + 4\phi_1\lambda f + \lambda^2(a_2 + a_1\lambda + \lambda^2) = 0 .$$

The above equation has the following solution:

$$f = -2\phi_1\lambda \pm \lambda\sqrt{4\phi_1^2 - (a_2 + a_1\lambda + \lambda^2)} ;$$

or, if α_1, α_2 are expressed by equation 15,

$$f = -2\phi_1\lambda \pm \lambda\sqrt{9k^2\phi_0^2 - \left[\lambda + \phi_0(3k-1) - \phi_2\right]^2} . \tag{17}$$

Equation 17, together with equations 13 and 16, provides the formal solution of the problem.

APPLICATION

If we consider the motion of the satellite as a disturbed Kepler motion, it is practical to recall that for $k=0$ the third of equations 7 transforms into the relation

$$r\dddot{r} + 3\dot{r}\ddot{r} + \frac{\dot{r}}{r^2} = 0 , \tag{18}$$

which can be integrated in a closed form. For, according to equation 8d, the disturbed motion is

$$r\dddot{r} + 3\dot{r}\ddot{r} + \frac{\dot{r}}{r^2} = \phi_3 - 4k\phi_1 .$$

But from equation 16,

$$\phi_0(\phi_3 - 4k\phi_1) = f + 2\phi_1(\phi_0 + \phi_2 - 2k\phi_0) ;$$

therefore

$$\phi_0\left(\dot{r}\ddot{r} + 3\dot{r}\ddot{r} + \frac{\dot{r}}{r^2}\right) = 2\phi_1(\phi_0 + \phi_2 - \lambda - 2k\phi_0) \pm \lambda\sqrt{9k^2\phi_0^2 - [\phi_0 + \phi_2 - \lambda - 3k\phi_0]^2} \tag{19}$$

The right side of this equation becomes zero for $k=0$, since, according to equation 12, $\phi_0+\phi_2-\lambda=x=6k\xi^2$ also contains k as a factor.

For the case in which the satellite's orbit can, to first approximation, be considered a Kepler ellipse, the right side of equation 19 is small, i.e., of the order of the perturbation. We then have, for $k=1$:

$$\left.\begin{aligned} \phi_0\left(r\dddot{r} + 3\dot{r}\ddot{r} + \frac{\dot{r}}{r^2}\right) &= 2\phi_1(\phi_2 - \phi_0 - \lambda) \pm \lambda\sqrt{9\phi_0^2 - (\phi_2 - 2\phi_0 - \lambda)^2} \\ \text{with} \qquad \lambda^2 &= \phi_0^2 + 2\phi_0(\phi_2 + 2\phi_4) - 4\phi_1^2 . \end{aligned}\right\} \tag{20}$$

The sign of the square root is determined from the initial conditions of the problem. μ itself is certainly positive if the motion of the satellite is direct, for, by definition,

$$\lambda = \phi_0 + \phi_2 - 6k\xi^2 = r^2 + 2(\xi\dot{\eta} - \eta\dot{\xi}) .$$

On the other hand, if, in the fixed xy-system, $x\dot{y}-y\dot{x}=r^2\dot{\psi}$ is the velocity at a given instant,

$$\xi\dot{\eta} - \eta\dot{\xi} = r^2(\dot{\psi} - 1) ,$$

so that

$$\lambda = r^2(2\dot{\psi} + 1).$$

Therefore λ is positive even for $\dot{\psi} > -\frac{1}{2}$.

The advantage of the differential equation (equation 20) is that it contains only one variable, so that the expression for the perturbation (on the right side) also depends only upon r, $\dot{r}$, $\ddot{r}$ and C. In the application of the methods of numerical integration, it suffices, therefore, to set up a single difference table. This step could be of particular importance when working with electronic computers. The relatively complicated structure of the disturbing function does not constitute an appreciable barrier for the computers.

It must also be mentioned that equation 20 gives the function $r(t)$ independently of the coordinate

system selected, and that therefore its solution is as valid in a fixed as in a revolving system. It is here that this solution differs from that which Hill gave for the system.[2] Hill's solution was limited to simple-periodic orbits in a revolving ξ, η system. Although the Hill experiments yield—of the known periodic inequalities—only the so-called "variation" which is periodic in the revolving system, a similar analysis of the differential equation (equation 20) would also yield the "large inequality" of the moon motion (largest periodic term of elliptical motion), the "evection," and some other features. Likewise, considering the secular perturbations, we find the apsidal motion, but it is inherent in the problem that the terms due to the solar parallax, the eccentricity and inclination of the earth's orbit do not appear.

Provided that we have succeeded in determining $r=r(t)$—either by numerical integration of equation 20 or by a theoretical formula (developed perhaps through a sufficiently converging series)—it is always possible to determine the corresponding coordinate $\phi(t)$ by means of a simple quadrature. By letting $\xi=r\cos\phi$ and $\eta=r\sin\phi$, Hill's equations (equation 1) take the form:

$$\left(\ddot{r} - r\dot{\phi}^2 - 2r\dot{\phi}\right)\cos\phi - \left(2\dot{r}\dot{\phi} + r\ddot{\phi} + 2\dot{r}\right)\sin\phi = \left(3 - \frac{1}{r^3}\right) r\cos\phi ,$$

$$\left(\ddot{r} - r\dot{\phi}^2 - 2r\dot{\phi}\right)\sin\phi + \left(2\dot{r}\dot{\phi} + r\ddot{\phi} + 2\dot{r}\right)\cos\phi = -\frac{1}{r^3} r\sin\phi$$

from which

$$\left.\begin{aligned} \ddot{r} - r\dot{\phi}^2 - 2r\dot{\phi} &= 3r\cos^2\phi - \frac{1}{r^2} , \\ 2\dot{r}\dot{\phi} + r\ddot{\phi} + 2\dot{r} &= -3r\sin\phi\cos\phi = -\frac{3}{2} r\sin 2\phi \end{aligned}\right\} \quad (21)$$

---

follow. Hill has shown from the second equation of equation 21 that $r=r(t)$ is found by quadrature if ϕ (t) is known, for it follows from

$$2\frac{\dot{r}}{r}(\dot{\phi}+1) = -\left(\ddot{\phi} + \frac{3}{2}\sin 2\phi\right)$$

that, with k constant,

$$r = \frac{k}{\sqrt{\dot{\phi}+1}} \exp -\frac{3}{4}\int \frac{\sin 2\phi}{\dot{\phi}+1}\, dt .$$

In the same manner we also find from the first equation of equation 21 that

$$\dot{\phi}^2 + 2\dot{\phi} = \frac{\ddot{r}}{r} + \frac{1}{r^3} - 3\cos^2\phi$$

or

$$\dot{\phi} + 1 = \sqrt{\frac{\ddot{r}}{r} + \frac{1}{r^3} + 1 - 3\cos^2\phi} , \quad (22)$$

from which it follows, by integration, that

$$t - t_0 = \int_{\phi_0}^{\phi} \frac{d\phi}{\sqrt{\frac{\ddot{r}}{r} + \frac{1}{r^3} + 1 - 3\cos^2\phi - 1}} . \quad (23)$$

The above relation yields $t=t(\phi)$ and, therefore $\phi=\phi(t)$.

If $\psi=\phi_0+nt=\phi_0+t$ is the true length of the satellite in a fixed coordinate system, then

$$\dot{\psi} = \dot{\phi} + 1 = \frac{\sqrt{p}}{r^2} ,$$

for undisturbed motion, if p denotes the parameter of the orbit ellipse; we also have

$$r^2\left(r\ddot{r} + \frac{1}{r}\right) = p ,$$

therefore

$$\dot{\phi} + 1 = \sqrt{\frac{\ddot{r}}{r} + \frac{1}{r^3}} . \quad (24)$$

This differential equation can therefore be con-

sidered as an approximation for equation 22 as long as the term $(1-3\cos^2\phi)$ in the radicand of the square root remains small by comparison with the other terms. For the earth's moon, which is of particular interest here, this condition is still satisfied, for, since the unit of length is so selected that the mean earth-sun distance is about 69, the average value of r (normally 390 times smaller) is about $\frac{3}{17}$; the expression $(\ddot{r}/r)+(1/r^3)$ is therefore of the order p/r^4 or $1/r^3$, or about 180, whereas the attachment term varies from -2 to $+1$.

INTRODUCTION OF A NEW INDEPENDENT VARIABLE

The form of the differential equation (equation 20) suggests a solving process which is a modification of that used in the undisturbed problem (equation 18). If a new variable q is introduced instead of time t by means of $\dot{q}=1/r$ and the initial condition $q(t_0)=0$ so that q disappears simultaneously and increases monotonically with $(t-t_0)$, we find that*

$$\left.\begin{aligned} \dot{r} &= r'\dot{q} = \frac{r'}{r}\,, \\ \ddot{r} &= \frac{r''}{r^2} - \frac{r'^2}{r^3}\,, \\ \dddot{r} &= \frac{r'''}{r^3} - 4\frac{r'r''}{r^4} + 3\frac{r'^3}{r^5}\,. \end{aligned}\right\} \qquad (25)$$

With the above equations, the undisturbed equation of motion (equation 18) assumes the form

$$\frac{1}{r^2}\left(r''' + \frac{1-r''}{r}\,r'\right) = 0\,. \qquad (26)$$

But since

$$\frac{d}{dq}\left(\frac{1-r''}{r}\right) = -\frac{1}{r}\left(r''' + \frac{1-r''}{r}\,r'\right) = 0\,, \qquad (27)$$

$$\alpha^2 = \frac{1-r''}{r}$$

is constant, and equation 26 goes over into the integrable form:

$$r''' + \alpha^2 r' = 0\,. \qquad (28)$$

On the other hand, in the disturbed problem (Equation 20) we have

$$r''' + \frac{1-r''}{r}\,r' = \phi(q) = \phi(r, r', r'', C)$$

or

$$r''' + \alpha^2 r' = \phi(q)\,, \qquad (29)$$

where $\phi(q)$ represents the expression for the perturbation (right side of equation 20) as a function of q. And because of equation 27,

$$\frac{d}{dq}(\alpha^2) = -\frac{\phi(q)}{r}$$

or

$$\alpha^2 = \alpha_0^2 - \int_0^q \frac{\phi(q)}{r(q)}\,dq\,. \qquad (30)$$

Therefore, if we set

$$g(q) = \phi + r'\int_0^q \frac{\phi}{r}\,dq\,, \qquad (31)$$

we get

$$r''' + \alpha_0^2 r' = g(q)\,, \qquad (32)$$

where $\alpha_0^2=(1-r_0'')/r_0$ is a constant which is known from the initial conditions.

If $g(q)$ is a sufficiently small perturbation in the interval $|q|<Q$, the integration method described in a companion paper[3] could be utilized. If, for $s=\alpha_0 q$, we set $c_0(s^2)=\cos s$, $c_1(s^2)=(\sin s)/s$, $c_2(s^2)=(1-\cos s)/s^2$ and $c_3(s^2)=(s-\sin s)/s^3, \cdots$, then the recurrence formula

$$1/n! = c_n + s^2 c_{n+2} \qquad (33)$$

and the differential equation

$$\frac{d}{dq}\left(c_{n+1}\,q^{n+1}\right) = c_n q^n\,. \qquad (34)$$

*The derivates of q are denoted by primes.

are applicable. Therefore the Taylor development

$$r(q) = r_0 + \frac{1}{1!} r_0' q + \frac{1}{2!} r_0'' q^2 + \frac{1}{3!} r_0''' q^3 + \cdots ,$$

can, by replacing the reciprocal factorials with the expressions given by equation 33, be written

$$r(q) = r_0 + r_0' \left(c_1 + s^2 c_3\right) q + r_0'' \left(c_2 + s^2 c_4\right) q^2 + \cdots .$$

If we again set $s^2 = \alpha^2 q^2$ and arrange according to powers of q, we have

$$r(q) = r_0 + r_0' c_1 q + r_0'' c_2 q^2 + \left(r_0''' + \alpha_0^2 r_0'\right) c_3 q^3 + \left(r_0^{iv} + \alpha_0^2 r_0''\right) c_4 q^4 + \cdots ,$$

a series which, because of equation 32, becomes

$$r(q) = r_0 + r_0' c_1 q + r_0'' c_2 q^2 + g_0 c_3 q^3 + g_0' c_4 q^4 + g_0'' c_5 q^5 + \cdots . \tag{35}$$

Now we set

$$r_0 q = z\tau , \tag{36}$$

where $\tau = k(t - t_0)$ denotes the intermediate time expressed in units of $1/k$ days; we introduce the expressions $\gamma_1, \gamma_2, \gamma_3, \cdots$ (which are invariable against coordinate transformations) through

$$r_0' = r_0^2 \gamma_1 , \quad r_0'' = r_0^3 \gamma_2 , \quad g_0 = r_0^4 \gamma_3 , \quad g_0' = r_0^5 \gamma_4 , \quad \cdots ,$$

and we set

$$\gamma_1 \tau = \eta_1 , \quad \gamma_2 \tau^2 = \eta_2 , \quad \cdots \quad \gamma_n \tau^n = \eta_n ,$$

Equation 35 now assumes the form

$$r(q) = r_0 \left[1 + c_1 \eta_1 z + c_2 \eta_2 z^2 + c_3 \eta_3 z^3 + \cdots\right] , \tag{37}$$

a series, the terms of which are small above the third order, and which, in practical cases and for moderate intermediate times, converges very rapidly. The association between τ and q is provided by the differential equation

$$d\tau = r\, dq = r_0 \left[1 + c_1 \eta_1 z + c_2 \eta_2 z^2 + \cdots\right] dq . \tag{38}$$

But

$$r_0 \int c_n \eta_n z^n dq = r_0 \gamma_n \int c_n (z\tau)^n dq = r_0^{n+1} \gamma_n \int c_n q^n dq ;$$

and therefore, according to equation 34,

$$r_0 \int c_n \eta_n z^n dq = r_0^{n+1} \gamma_n c_{n+1} q^{n+1} = \gamma_n c_{n+1} z^{n+1} \tau^{n+1} = \tau \eta_n c_{n+1} z^{n+1} .$$

If we divide the integral of equation 38 by τ, we obtain the *main equation*

$$1 = z + c_2 \eta_1 z^2 + c_3 \eta_2 z^3 + c_4 \eta_3 z^4 + \cdots . \tag{39}$$

For undisturbed motion ($g = 0$), it limits itself to the first three terms, since $\eta_3, \eta_4, \cdots$ disappear.

The c-functions

$$c_n(\lambda^2) = \frac{1}{n!} - \frac{\lambda^2}{(n+2)!} + \frac{\lambda^4}{(n+4)!} - \cdots$$

have the always-real argument

$$\lambda^2 = (\alpha_0 q)^2 = \alpha_0^2 \frac{z^2 \tau^2}{r_0^2} ;$$

or, since

$$\alpha_0^2 = \frac{1 - r_0''}{r_0} = \frac{1 - r_0^3 \gamma_2}{r_0} r_0^2 \left(\frac{1}{r_0^3} - \gamma_2 \right) ,$$

$$\lambda^2 = \left(\frac{1}{r_0^3} - \gamma_2 \right) \tau^2 z^2 .$$

CLOSING COMMENTS AND OUTLOOK

The aforementioned integration method replaces the numerical integration method of Hill's differential equation (equation 1) for the case where we can, as a first approximation, consider the satellite motion as a Keplerian ellipse with any desired eccentricity. This method permits the determination of $r(t)$ directly through iteration—at least under some conditions, for a certain fairly extended time interval in the vicinity of the initial time t_0. The iteration process itself is limited to the solution of the transcendental *main equation* (equation 39), since the quantities g_0, g_0', $\cdots$ can be derived from the initial conditions.

Here we have attempted to demonstrate this method on a problem—Hill's moon problem—which perhaps has little practical importance, but is simple enough that its solution may be written explicitly and understandably. Thus it serves as an example for other, more difficult problems, which can be handled with the same general principles and solved in the same manner. This method can be used to attack certain special cases of the *problème restreint*, in which conic-section orbits can be considered good approximate solutions. For example, we could consider the case of the sufficiently close orbit of an infinitesimal body around one of two finite masses; i.e., either the actual satellite problem in which the massless body moves around the lesser mass (planet) or the problem of disturbed planetoid orbits, when the motion occurs about the greater mass (sun) and the distance from the disturbing planet remains sufficiently large.

In these cases it is also possible to apply the aforementioned method rationally, since we are again concerned with a problem of motion of the fourth order, which can be reduced to the third order by applying the Jacobian integral. But here the difficulties are considerably greater, although not insurmountable. The elimination process described becomes much more complicated because the quantities to be eliminated (ξ, η, $\dot{\xi}$, $\dot{\eta}$) appear not only in the distance and velocity of the body in reference to its central mass, but also in the distance and velocity in reference to the disturbing mass. In Hill's problem, the last step of this process leads to the elimination of λ from the two algebraic equations (equations 13 and 14) which are of the second and fourth degrees. The resultant equation (equation 15), which solves the problem, is therefore obtained in the form of a six-row Sylvester determinant which is set equal to zero. The same process applied to the *problème restreint* leads to a correspondingly to-be-determined quantity λ that is to be eliminated from the two algebraic equations of the 10th and 14th degree, so that the final solution appears as a 24-row Sylvester determinant which is equated to zero, the strict solution of which would be hopelessly complicated. However, as long as we restrict ourselves to those cases in which the motion of the satellite or planetoid can be viewed as an only slightly disturbed Kepler orbit, the elimination can be accomplished with the help of a rapidly converging approximation process. The demonstration of this must be reserved for another treatise.

REFERENCES

1. Hill, G. W., "Researches in the Lunar Theory," in *The Collected Works of George William Hill, Vol. I*, Washington, D.C.: Carnegie Institution, 1905.
2. Hill, G. W., "Researches in the Lunar Theory," in *The Collected Works of George William Hill, Vol. I*, Washington, D.C.: Carnegie Institution, 1905.
3. Stumpff, K., "Calculation of Ephemerides from Initial Values," NASA Technical Note D-1415, 1962.

REMARKS ON HILL'S LUNAR THEORY, PART II

KARL STUMPFF

Goddard Space Flight Center

In 1879 G. W. Hill, using his differential equation system, found a series of direct lunar orbits, which represent a simplification of the restricted three-body problem (neglecting solar parallax and eccentricity). If T is the period of revolution time, and $T=2\pi m$ is fixed, the initial conditions, $\xi_0(m)$, $\dot{\eta}_0(m)$, and $(\dot{\xi}_0=\eta_0=0)$, are dependent upon m and give periodic solutions. Hill obtained these functions in the form of power series, with respect to m, which converges rapidly for small distances of the satellite; and he gave their first coefficients.

In the present discussion, Hill's coefficients are successively determined from linear equations and are in the form of closed trigonometric expressions. Although this method—which does not extend beyond the third order—is only an interesting attempt at achieving the same goal via a different mathematical route, it does not extend as far as Hill's method (which may be carried out to the fifth or sixth order without difficulty).

INTRODUCTION

In his famous writing of 1878, G. W. Hill[1] found a series of direct lunar orbits as solutions of his differential equation system. This series constitutes a simplification of the *problème restreint* (by discounting the solar parallax). Hill found that if T is the (synodic) circuit time of the satellite and $T=2\pi m$ is fixed, the initial conditions are $\xi_0(m)$, $\dot{\eta}_0(m)$, $(\dot{\xi}_0=\eta_0=0)$, and are dependent upon m as a parameter and give such periodic solutions. He obtained these functions in a power series with respect to m, which converge rapidly for small distances of the satellite (short circuit times, small m), and gave their first coefficients. In the following examination, another method will be presented which leads to the same goal, but which permits these coefficients to appear as finite trigonometric expressions.

HILL'S DIFFERENTIAL EQUATIONS

Hill's differential equations

$$\left.\begin{aligned} \ddot{\xi} - 2\dot{\eta} &= \xi\left(3 - \frac{1}{r^3}\right), \\ \ddot{\eta} + 2\dot{\xi} &= -\frac{\eta}{r^3}, \end{aligned}\right\} \qquad (1)$$

where $r^2=\xi^2+\eta^2$, are based upon the rectangular coordinates of a massless satellite in reference to its central body (planet) and in a coordinate system which revolves at the same angular velocity, $n=1$, as the planet which orbits the sun (fig. 1). The location of the sun S is therefore assumed to be in the direction of the negative ξ-axis, at any desired distance, so that the parallax of the sun can be discounted. Under these circumstances, equations 1 apply for the satellite S. Periodic solutions can be expected if we let the satellite start from a point $(\xi_0, 0)$ on the ξ-axis with a velocity $(0, \dot{\eta}_0)$, whereby the initial (positive) velocity must occur perpendicular to the ξ-axis and $\dot{\eta}_0$ will depend to a certain extent upon ξ_0. Every periodic solution of this sort is symmetric to both axes, since equations 1 do not change when ξ is

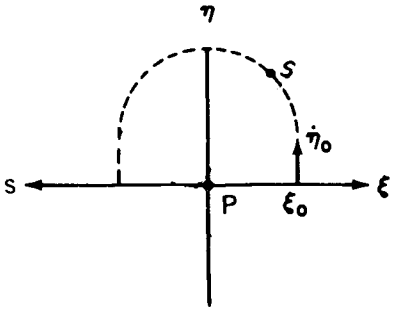


FIGURE 1.—Geometric representation of a massless satellite S of a planet P.

exchanged with $-\xi$ and η with $-\eta$. If $T=2\pi m$ is the circuit time in such a periodic orbit, then the total of these orbits also can be characterized by determining $\xi_0=\xi_0(m)$, $\dot{\eta}_0=\dot{\eta}_0(m)$ as functions of the parameter m.

If $\tau=0$ is the moment at which the satellite crosses the positive ξ-axis at right angles, then the periodic motion has the form

$$\left.\begin{aligned} \xi(\tau) &= A_1 \cos \nu\tau + A_3 \cos 3\nu\tau + A_5 \cos 5\nu\tau + \cdots , \\ \eta(\tau) &= B_1 \sin \nu\tau + B_3 \sin 3\nu\tau + B_5 \sin 5\nu\tau + \cdots , \end{aligned}\right\} \tag{2}$$

where $\nu=1/m=2\pi/T$. For circuit times as brief as described ($m\to 0$)—hence for very close orbits—the perturbations by the sun become as small as desired. The above equations (Equation 2) now change into

$$\xi(\tau) = a \cos \nu\tau, \quad \eta(\tau) = a \sin \nu\tau , \tag{3}$$

or into the form of uniform circular motion, for circles are the only undisturbed orbits which are symmetric to both axes. If we differentiate equation 3 any number of times, we find, for this boundary case and for $\tau=0$,

$$m^{2n} \xi_0^{(2n)} = (-1)^n \xi_0 ,$$

$$m^{2n-1} \eta_0^{2n-1} = (-1)^{n-1} \xi_0 .$$

The above equations mean that, for $m\to 0$, the quantities

$$\left.\begin{aligned} \alpha_2 &= -m^2 \frac{\ddot{\xi}_0}{\xi_0} \to 1 , & \beta_1 &= +m \frac{\dot{\eta}_0}{\xi_0} \to 1 , \\ \alpha_4 &= +m^4 \frac{\xi_0^{(4)}}{\xi_0} \to 1 , & \beta_3 &= -m^3 \frac{\dddot{\eta}_0}{\xi_0} \to 1 , \\ \alpha_6 &= -m^6 \frac{\xi_0^{(6)}}{\xi_0} \to 1 , & \beta_5 &= +m^5 \frac{\eta_0^{(5)}}{\xi_0} \to 1 , \end{aligned}\right\} \tag{4}$$

all approach unity.

If for the sake of abbreviation we set $1/r^3=\mu$, then the differential equations (equation 1) can be

written:

$$\ddot{\xi} = 2\dot{\eta} - \xi(\mu - 3) \; ; \quad \text{and} \quad \ddot{\eta} = -2\dot{\xi} - \eta\mu \; .$$

If we differentiate these equations any number of times and if we consider that uneven derivatives of ξ and μ and even derivatives of η disappear for $\tau = 0$, we then obtain the system:

$$\frac{-m^2}{\xi_0} \qquad \ddot{\xi}_0 = 2\dot{\eta}_0 - \xi_0(\mu_0 - 3) \; ,$$

$$\frac{+m^4}{\xi_0} \qquad \xi_0^{(4)} = 2\dddot{\eta}_0 - \left[\ddot{\xi}_0(\mu_0 - 3) + \xi_0 \ddot{\mu}_0\right] \; ,$$

$$\frac{-m^6}{\xi_0} \qquad \xi_0^{(6)} = 2\eta_0^{(5)} - \left[\xi_0^{(4)}(\mu_0 - 3) + 6\ddot{\xi}_0 \ddot{\mu}_0 + \xi_0 \mu_0^{(4)}\right]$$

$$\frac{-m^3}{\xi_0} \qquad \dddot{\eta}_0 = -2\ddot{\xi}_0 - \dot{\eta}_0 \mu \; ,$$

$$\frac{+m^5}{\xi_0} \qquad \eta_0^{(5)} = -2\xi_0^{(4)} - \left(\dddot{\eta}_0 \mu + 3\dot{\eta}_0 \ddot{\mu}\right) \; ,$$

$$\frac{-m^7}{\xi_0} \qquad \eta_0^{(7)} = -2\xi_0^{(6)} - \left(\eta_0^{(5)} \mu + 10\, \dddot{\eta}_0 \ddot{\mu} + 5\dot{\eta}_0 \mu^{(4)}\right) \; ,$$

If these equation are multiplied by the factors to their left and if equation 4 is substituted they become:

$$\left.\begin{aligned}
\alpha_2 &= -2m\beta_1 - 3m^2 + \phi \; , & \beta_1 &= \psi \; , \\
\alpha_4 &= -2m\beta_3 - 3m^2 \alpha_2 + \phi\nu_2 \; , & \beta_3 &= -2m\alpha_2 + \phi\nu_1 \; , \\
\alpha_6 &= -2m\beta_5 - 3m^2 \alpha_4 + \phi\nu_4 \; , & \beta_5 &= -2m\alpha_4 + \phi\nu_3 \; , \\
\alpha_8 &= -2m\beta_7 - 3m^2 \alpha_6 + \phi\nu_6 \; , & \beta_7 &= -2m\alpha_6 + \phi\nu_5 \; ,
\end{aligned}\right\} \qquad (5)$$

where

$$\phi = m^2 \mu_0 \quad \text{and} \quad \psi = \beta_1 = m\frac{\dot{\eta}_0}{\xi_0} \; ; \qquad (6)$$

and, with

$$\mu_2 = -m^2 \frac{\ddot{\mu}_0}{\mu_0} \; , \quad \mu_4 = +m^4 \frac{\mu_0^{(4)}}{\mu_0} \; , \quad \mu_6 = -m^6 \frac{\mu_0^{(6)}}{\mu_0} \; , \quad \cdots \; , \qquad (7)$$

$$\left.\begin{aligned} \nu_1 &= \beta_1 , & \nu_2 &= \alpha_2 + \mu_2 , \\ \nu_3 &= \beta_3 + 3\mu_2\beta_1 , & \nu_4 &= \alpha_4 + 6\mu_2\alpha_2 + \mu_4 , \\ \nu_5 &= \beta_5 + 10\mu_2\beta_3 + 5\mu_4\beta_1 , & \nu_6 &= \alpha_6 + 15\mu_2\alpha_4 + 15\mu_4\alpha_2 + \mu_6 . \end{aligned}\right\} \tag{8}$$

If the quantities $\mu_2, \mu_4, \cdots$ are eliminated from equation 8 by forming the even derivatives of μ and substituting their values for $\tau=0$ into equation 7, the following system of recurrence formulas is obtained for ν_i:

$$\left.\begin{aligned} \nu_1 &= \beta_1 , \\ \nu_3 &= \beta_3 - 3\alpha_2\nu_1 + 3\beta_1\nu_2 , \\ \nu_5 &= \beta_5 - 5\alpha_4\nu_1 + 10\beta_3\nu_2 - 10\alpha_2\nu_3 + 5\beta_1\nu_4 , \\ \nu_2 &= -(1+1)\alpha_2 + (1+2)\beta_1\nu_1 , \\ \nu_4 &= -(1+1)\alpha_4 + (3+4)\beta_3\nu_1 - (3+6)\alpha_2\nu_2 + (1+4)\beta_1\nu_3 , \\ \nu_6 &= -(1+1)\alpha_6 + (5+6)\beta_5\nu_1 - (10+15)\alpha_4\nu_2 + (10+20)\beta_3\nu_3 - (5+15)\alpha_2\nu_4 + (1+6)\beta_1\nu_5 , \end{aligned}\right\} \tag{9}$$

In these formulas we again recognize the binomial coefficients of uneven order in the coefficients of ν_i with uneven indices, while the coefficients of ν_i with even indices are made up of the binomial coefficients of uneven and subsequent even orders. In all cases, the signs are alternating and the sum of the coefficients is unity. However, this means that all ν_i, just as the α_i and β_i, approach unity for $m\to 0$. The same also applies for ϕ and ψ, as can be seen from equation 5. With the help of equations 5 and 9, α_i and β_i can be determined one after another, in the order of increasing indices ($\beta_1, \alpha_2, \beta_3, \alpha_4, \cdots$), as functions of m, ϕ, ψ.

The orbit of a satellite which originates from $(\xi_0, 0)$ at time $\tau=0$ with the velocity $(0, \dot{\eta}_0)$ will be periodic if it reaches the point $(-\xi_0, 0)$ at $\tau=\frac{1}{2}T=\pi m$. Therefore the periodicity conditions can be written:

$$\xi(\pi m) = -\xi_0 , \quad \eta(\pi m) = 0 , \tag{10}$$

or, if we set up the Taylor series,

$$\xi(\pi m) = -\xi_0 = \xi_0 + \frac{(\pi m)^2}{2!}\ddot{\xi}_0 + \frac{(\pi m)^4}{4!}\xi_0^{(4)} + \cdots ,$$

$$\eta(\pi m) = 0 = \pi m\dot{\eta}_0 + \frac{(\pi m)^3}{3!}\dddot{\eta}_0 + \frac{(\pi m)^5}{5!}\eta_0^{(5)} + \cdots .$$

If we divide by ξ_0 and substitute the quantities of equation 4, we can then write

$$\left.\begin{aligned} f(m) &= 0 = 2 - \alpha_2 \frac{\pi^2}{2!} + \alpha_4 \frac{\pi^4}{4!} - \alpha_6 \frac{\pi^6}{6!} + \cdots \\ g(m) &= 0 = \beta_1 \pi - \beta_3 \frac{\pi^3}{3!} + \beta_5 \frac{\pi^5}{5!} - \cdots . \end{aligned}\right\} \tag{11}$$

On the other hand,

$$\left.\begin{aligned} f(m) &= f_0 + f_0' m + f_0'' \frac{m^2}{2!} + f_0''' \frac{m^3}{3!} + \cdots \equiv 0 , \\ g(m) &= g_0 + g_0' m + g_0'' \frac{m^2}{2!} + g_0''' \frac{m^3}{3!} + \cdots \equiv 0 , \end{aligned}\right\} \tag{12}$$

if the index 0 is based upon $m=0$ and the derivatives according to m indicated by primes. The equations 12 are identities, since they are fulfilled for any given value of m. Thus the requirement of periodic orbits is given by

$$\left.\begin{aligned} f_0 &= f_0' = f_0'' = f_0''' = \cdots = 0 , \\ g_0 &= g_0' = g_0'' = g_0''' = \cdots = 0 . \end{aligned}\right\} \tag{13}$$

In general, then, if n is any positive whole number or zero and if $\alpha_i^{(n)}$, $\beta_i^{(n)}$ are understood to be the nth derivatives of the quantities α_i, β_i with respect to m for $m=0$, then

$$\left.\begin{aligned} f_0 &= 2 - \frac{\pi^2}{2!} + \frac{\pi^4}{4!} - \frac{\pi^6}{6!} + \cdots = 1 + \cos\pi = 0 , \\ g_0 &= \pi - \frac{\pi^3}{3!} + \frac{\pi^5}{5!} - \frac{\pi^7}{7!} + \cdots = \sin\pi = 0 , \end{aligned}\right\} \tag{14}$$

since all α_i, β_i assume the value 1 for $m \to 0$, and furthermore

$$\left.\begin{aligned} f_0^{(n)} &= 0 = -\alpha_2^{(n)} \frac{\pi^2}{2!} + \alpha_4^{(n)} \frac{\pi^4}{4!} - \alpha_6^{(n)} \frac{\pi^6}{6!} + \cdots \\ g_0^{(n)} &= 0 = \beta_1^{(n)} \pi - \beta_3^{(n)} \frac{\pi^3}{3!} + \beta_5^{(n)} \frac{\pi^5}{5!} - \cdots . \end{aligned}\right\} \tag{15}$$

The equations 14 are identically fulfilled and only show that the arrangements were correct. For additional computations we shall make an analysis of equations 15 for $n=1$ explicitly; i.e.,

$$\left.\begin{aligned} f_0' &= 0 = -\alpha_2' \frac{\pi^2}{2!} + \alpha_4' \frac{\pi^4}{4!} - \alpha_6' \frac{\pi^6}{6!} + \cdots \\ g_0' &= 0 = \beta_1' \pi - \beta_3' \frac{\pi^3}{3!} + \beta_5' \frac{\pi^5}{5!} - \cdots . \end{aligned}\right\} \tag{16}$$

We have shown above that it is possible to represent the quantities α_i, β_i as functions of m, ϕ and ψ.

Since ϕ and ψ assume the value 1 for $m=0$, and if we set

$$\left.\begin{aligned} \phi(m) &= 1 + a_1 m + a_2 \frac{m^2}{2!} + a_3 \frac{m^3}{3!} + \cdots \\ \psi(m) &= 1 + b_1 m + b_2 \frac{m^2}{2!} + b_3 \frac{m^3}{3!} + \cdots , \end{aligned}\right\} \tag{17}$$

it follows, for $m=0$, that

$$\phi_0^{(n)} = a_n , \quad \psi_0^{(n)} = b_n . \tag{18}$$

From equations 5 and 9 we obtain, one after another, the equations:

$$\beta_1' = \psi' ,$$

$$\alpha_2' = -2\beta_1 - 2m\beta_1' - 6m + \phi' ,$$

$$\beta_3' = -2\alpha_2 - 2m\alpha_2' + \phi'\beta_1 + \phi\beta_1' ,$$

$$\alpha_4' = -2\beta_3 - 2m\beta_3' - 6m\alpha_2 - 3m^2\alpha_2' + \phi'\nu_2 + \phi\left(-2\alpha_2' + 6\beta_1\beta_1'\right), \text{ etc.}$$

If, in these and the following formulas, we set $m=0$ and consider that the α_i, β_i, ν_i, ϕ, and ψ go towards 1 and the derivations ϕ', ψ' towards a_1, b_1, then, after some brief calculations, we get

$$\alpha_2' \to a_1 - 2 , \qquad \beta_1' \to b_1 ,$$

$$\alpha_4' \to 6b_1 - (a_1 - 2) , \qquad \beta_3' \to b_1 + (a_1 - 2) ,$$

$$\alpha_6' \to 48b_1 - 21(a_1 - 2) , \qquad \beta_5' \to 19b_1 - 7(a_1 - 2) ,$$

$$\alpha_8' \to 234b_1 - 113(a_1 - 2) , \qquad \beta_7' \to 109b_1 - 51(a_1 - 2) .$$

If we substitute these expressions into equation 16, we get two linear equations:

$$\left.\begin{aligned} Ab_1 - B(a_1 - 2) &= 0 , \\ Cb_1 - D(a_1 - 2) &= 0 , \end{aligned}\right\} \tag{19}$$

the coefficients of which are represented by the π-series:

$$A = 6\frac{\pi^4}{4!} - 48\frac{\pi^6}{6!} + 234\frac{\pi^8}{8!} - \cdots ,$$

$$B = \frac{\pi^2}{2!} + \frac{\pi^4}{4!} - 21\frac{\pi^6}{6!} + 113\frac{\pi^8}{8!} - \cdots ,$$

$$C = \pi - \frac{\pi^3}{3!} + 19\frac{\pi^5}{5!} - 109\frac{\pi^7}{7!} + \cdots ,$$

$$D = \frac{\pi^3}{3!} + 7\frac{\pi^5}{5!} - 51\frac{\pi^7}{7!} + \cdots .$$

This series can be easily summed, for from

$$\cos\pi = 1 - \frac{\pi^2}{2!} + \frac{\pi^4}{4!} - \frac{\pi^6}{6!} + \frac{\pi^8}{8!} - \cdots ,$$

$$\frac{1}{2}\cos 2\pi = \frac{1}{2} - 2\frac{\pi^2}{2!} + 8\frac{\pi^4}{4!} - 32\frac{\pi^6}{6!} + 128\frac{\pi^8}{8!} - \cdots ,$$

$$2\pi\sin\pi = 4\frac{\pi^2}{2!} - 8\frac{\pi^4}{4!} + 12\frac{\pi^6}{6!} - 16\frac{\pi^8}{8!} + \cdots$$

we obtain by summation:

$$\cos\pi + \frac{1}{2}\cos 2\pi + 2\pi\sin\pi = \frac{3}{2} + \left(\frac{\pi^2}{2!} + \frac{\pi^4}{4!} - 21\frac{\pi^6}{6!} + 113\frac{\pi^8}{8!} - \cdots\right)$$

$$= \frac{3}{2} + B = -2 .$$

Similarly we represent A, C, D by closed expressions: specifically,

$$\left.\begin{aligned} A &= -3 + 2\cos\pi + \cos 2\pi + 3\pi\sin\pi = -4 , \\ B &= -\frac{3}{2} + \cos\pi + \frac{1}{2}\cos 2\pi + 2\pi\sin\pi = -2 , \\ C &= 2\sin\pi + \sin 2\pi - 3\pi\cos\pi = 3\pi , \\ D &= \sin\pi + \frac{1}{2}\sin 2\pi - 2\pi\cos\pi = 2\pi . \end{aligned}\right\} \quad (20)$$

Since the determinant of the homogeneous system (equation 19), $BC-AD=2\pi$, differs from zero, it follows that

$$a_1 = 2, \quad b_1 = 0 . \quad (21)$$

After a_1, b_1 are known, we can determine a_2, b_2 in a corresponding manner from the pair of equations $f_0''=0$, $g_0''=0$. From these a pair of linear equations results:

$$\left.\begin{aligned} Ab_2 - Ba_2 &= X_2 , \\ Cb_2 - Da_2 &= Y_2 , \end{aligned}\right\} \quad (22)$$

in which A, B, C, D are defined as in equation 20, and in which X_2, Y_2 represent new π-series. The coefficients of this π-series can be determined numerically by using equation 21. With the values (equation

20) for $A, \cdots D$, we find the general expressions

$$\left.\begin{aligned} Ab_n - Ba_n &= X_n , \qquad a_n = -\frac{3}{2} X_n - \frac{2}{\pi} Y_n , \\ Cb_n - Da_n &= Y_n , \qquad b_n = -X_n - \frac{1}{\pi} Y_n , \end{aligned}\right\} \tag{23}$$

where X_n, Y_n are known, if all a_i, b_i up to $i=n-1$ are present numerically.

By carrying out this relatively easy analysis to the order $n=3$, we obtain:

$$\left.\begin{aligned} X_1 &= 3 - 2\cos\pi - \cos 2\pi - 4\pi \sin\pi = -2B = +4 , \\ X_2 &= -3 + \frac{19}{8}\cos\pi + \cos 2\pi - \frac{3}{8}\cos 3\pi - \frac{3}{2}\pi \sin\pi = -4 , \\ X_3 &= \frac{15}{2} - 2\cos\pi - \frac{5}{2}\cos 2\pi - 3\cos 3\pi - 36\pi \sin\pi = +10 , \\ Y_1 &= -2\sin\pi - \sin 2\pi + 4\pi\cos\pi = -2D = -4\pi , \\ Y_2 &= -\frac{19}{8}\sin\pi + \sin 2\pi - \frac{3}{8}\sin 3\pi + \frac{3}{2}\cos\pi = \frac{3}{2}\pi , \\ Y_3 &= -22\sin\pi - \frac{5}{2}\sin 2\pi - 3\sin 3\pi + 36\pi\cos\pi = -36\pi . \end{aligned}\right\} \tag{24}$$

The first coefficients of the development (equation 17) are therefore known, and we obtain

$$\phi(m) = 1 + 2m + 9\frac{m^2}{2!} + 57\frac{m^3}{3!} + \cdots ,$$

and

$$\psi(m) = 1 + \frac{11}{2}\frac{m^2}{2!} + 26\frac{m^3}{3!} + \cdots .$$

But according to equation 6, we have

$$\phi = \frac{m^2}{r_0^3} = \frac{m^2}{\xi_0^3} , \qquad \psi = m\frac{\dot{\eta}_0}{\xi_0} ,$$

so that we can also calculate ϕ and ψ from Hill's power series for ξ_0 and $\dot{\eta}_0$. This calculation yields:

$$\phi(m) = 1 + 2m + 9\frac{m^2}{2!} + 57\frac{m^3}{3!} + \frac{1649}{4}\frac{m^4}{4!} + \frac{1235611}{324}\frac{m^5}{5!} + \cdots , \tag{25a}$$

$$\psi(m) = 1 + \frac{11}{2}\frac{m^2}{2!} + 26\frac{m^3}{3!} + \frac{3857}{24}\frac{m^4}{4!} + \frac{23153}{18}\frac{m^5}{5!} + \cdots , \tag{25b}$$

and completely verifies our analysis.

CONCLUDING REMARKS

The foregoing method of obtaining Hill's coefficients is not intended to be more than an interesting attempt to achieve the same goal via a completely different mathematical route. This attempt is interesting for two reasons: (1) Because the coefficients of the series (equations 25a and 25b) are determined successively from linear equations (equation 23); and (2) Because they are in the form of closed trigonometric expressions. It must not be overlooked, however, that this method—elegant though it may be—does not lead as far as Hill's, the developments of which can be carried out to the fifth or sixth order without difficulty. Here we could hardly pass beyond the third order—at least not with the method used, which is, however, probably not the only possible one. The reason for this is as follows:

1. The method of determining the coefficients of the trigonometric formulas (equation 24) may be demonstrated by using the example of X_2, for which we obtain the π-series

$$X_2 = -6\frac{\pi^2}{2!} - 6\frac{\pi^4}{4!} + 198\frac{\pi^6}{6!} - 2190\frac{\pi^8}{8!} + \cdots . \tag{26}$$

Then the arrangement

$$\left.\begin{aligned} X_2 &= a + b\cos\pi + c\cos 2\pi + d\cos 3\pi + e\pi\sin\pi \\ &= a + b\left(1 - \frac{\pi^2}{2!} + \frac{\pi^4}{4!} - \cdots\right) \\ &\quad + c\left(1 - 4\frac{\pi^2}{2!} + 16\frac{\pi^4}{4!} - \cdots\right) \\ &\quad + d\left(1 - 9\frac{\pi^2}{2!} + 81\frac{\pi^4}{4!} - \cdots\right) \\ &\quad + e\left(2\frac{\pi^2}{2!} - 4\frac{\pi^4}{4!} - \cdots\right) \end{aligned}\right\} \tag{27}$$

leads to the equations:

$$\left.\begin{aligned} a + b + c + d &= 0 , \\ -b - 4c - 9d + 2e &= -6 , \\ b + 16c + 81d - 4e &= -6 , \\ -b - 64c - 729d + 6e &= 198 , \\ b + 256c + 6561d - 8e &= -2190 , \end{aligned}\right\} \tag{28}$$

the number of which can be expanded as far as the coefficients of the π-series are known. Only if the first five of these equations yield the coefficients of the trigonometric arrangement (equation 27), will this

formula solve the problem exactly; however, when the coefficients—$a, \cdots e$—calculated by means of it also fulfill the other equations (equation 28), a demand which could only be satisfied if it were possible to find a law (and then to prove it) by which the terms on the right side of equation 28—the coefficients of equation 26—would increase.

In the case of the developments (equation 24), it was possible to check the validity of the formula with at least *one* of the supernumerary equations and to thereby prove it. Furthermore, in this case, it is not difficult to find the law of the progression of the coefficients. The agreement with the results of Hill also vouch for their correctness.

2. The difficulties of extending this method begin with the determination of X_4, Y_4, for the numerical values of the coefficients of the π-series (equation 26) increase extraordinarily, and because until now it has remained uncertain which formula (equation 27) would achieve the right goal. A trial of the formula

$$X_4 = a + b\cos\pi + \cos 2\pi + d\cos 3\pi + e\cos 4\pi + f\pi\sin\pi$$

and a corresponding one for Y_4 has not succeeded, probably because the formula requires still more terms—at least one with $\cos 5\pi$.

RECOMMENDATIONS

Further experiments with this mathematical problem could be useful and interesting, and we shall conclude this report with some suggestions:

1. The expressions of equation 24 can be given in a simpler form. For instance, we could write

$$X_1 = 2(1-\cos\pi) + 2\sin^2\pi - 4\pi\sin\pi,$$

and by the substitution of the "c-functions" known from the two-body motion,[2]

$$c_0 = \cos\lambda, \quad c_1 = (\sin\lambda)/\lambda, \quad c_2 = (1-\cos\lambda)/\lambda^2, \quad c_3 = (\lambda-\sin\lambda)/\lambda^3, \cdots,$$

we then obtain, for $\lambda = \pi$:

$$X_1 = \pi^2\left[2\left(c_2 + c_1^2\right) - 4c_1\right],$$

$$X_2 = \pi^2\left[-2\left(c_2 + c_1^2\right) + \frac{3}{2}c_1\left(c_0c_1 - 1\right)\right],$$

$$X_3 = \pi^2\left[5\left(c_2 + c_1^2\right) + 12c_1\left(c_0c_1 - 3\right)\right],$$

and

$$Y_1 = \pi^3\left[2\left(c_3 - c_2\right) + 2c_0c_3\right],$$

$$Y_2 = \pi^3\left[\frac{3}{2}\left(c_3 - c_2\right) + c_1\left(\frac{3}{2}c_1^2 - 2c_2\right)\right],$$

$$Y_3 = \pi^3\left[36\left(c_3 - c_2\right) + c_1\left(12c_1^2 + 5c_2\right)\right].$$

2. Noticeable and displeasing is the fact that the expressions of equation 24 contain a rather large number of disappearing terms, which are necessarily created by the algorithm which leads to them. It should be possible to modify the process in such a way that these superfluous and complicating terms can be avoided at the beginning. Here is a mathematical problem that is not only challenging, but its solution would be very useful in celestial mechanics wherever the problem of seeking initial conditions for periodic orbits exists.

REFERENCES

1. HILL, G. W., "Researches in the Lunar Theory," in *The Collected Works of George William Hill, Vol. I*, Washington, D.C.: Carnegie Institution, 1905.
2. STUMPFF, K., "Calculation of Ephemerides from Initial Values," NASA Technical Note D–1415, 1962.

THE PERTURBATIONS OF A 24-HOUR NEAR EQUATORIAL SATELLITE DUE TO AN EARTH GRAVITY POTENTIAL THROUGH 4TH ORDER

C. A. WAGNER

Goddard Space Flight Center

This report extends previous investigations of 24 hours near equatorial earth satellites[1–3] by considering the motion of such satellites in an earth gravity field to 4th order. The three coupled second order linear differential equations of initial drift from a 24-hour equatorial circular reference orbit, are presented. This linear system is analyzed for "stable regions" in the field within which drift motion is self limiting. A somewhat simplified form of the linear equations are integrated to give a general solution which describes the overall features of the motion away from injection.

NOMENCLATURE

$\mathbf{F}$	The earth's gravity field.
m	A point mass.
V_E	The earth's gravity potential field.
μ_E	The earth's gaussian gravitational constant.
R_0 or $R_\circ$	The mean equatorial radius of the earth ellipsoid.
n, m	When used in the earth potential function; indicates the harmonic of order n and power m.
r_s	A nominal or calculated orbit radius at injection for a near synchronous earth satellite.
$\Delta r, \Delta\lambda, \Delta\phi$	Perturbation coordinates in radius, geographic longitude and latitude from a reference equatorial orbit which is circular, having the near synchronous radius r_s and a period of exactly 1 sidereal day. These coordinates measure the drift of a near synchronous earth satellite from a point in space at a fixed r_s from the c.m. of the earth, moving along the equator at the earth's rotation rate so as to maintain a fixed geographical longitude at all times.
r_1, λ_1, ϕ_1	Dimensionless perturbation coordinates. $r_1 = \Delta r/r_s$, $\lambda_1 = \Delta\lambda$ (rad.), $\phi_1 = \Delta\phi$ (rad.).
T	Dimensionless time variable. $T =$ wt, where $w =$ the earth's sidereal rotation rate, and t is real time.
$\dot{\theta}_E, w$	The earth's "constant" sidereal rotation rate.
d	The total differential operator.
$\gamma_{nm}, J_{nm}, B_{nm}, \theta_{nm}$	Constants of the earth's gravity potential: Defined as they are used in the text of the report.
$a, b, c, d, e, f, g, h, i, j, k, l, m, n$	Perturbation constants of the coupled linearized perturbation equations of motion with respect to the synchronous circular orbit of radius r_s.

s	A differential operator. Thus, $s^0(\ ) = (\ )$; $s^1(\ ) = d(\ )/dT$, $s^2(\ ) = d^2(\ )/dT$, etc. And, $s_0{}^0(\ ) = (\ )$ at time $T=0$; etc.
$A, B, C, D, E,$ A_1, A_2, A_3	Perturbation constants of the uncoupled perturbation equations of motion of the near synchronous satellite of the earth.
$(\cdot), (\cdot\cdot)$	Differentiation with respect to real time t, prior to (14), and differentiation with respect to dimensionless time T after (14).
t	Real time, from a zero at the point of injection of the satellite into it's near synchronous orbit.
r, θ, ϕ	Spherical coordinates of geocentric radius, inertial longitude from the vernal equinox, and geocentric latitude from the earth's equator, locating the near synchronous satellite m in inertial space.
r, λ, ϕ	Spherical coordinates of the near synchronous satellite; geocentric radius; geographic longitude with respect to the Greenwich meridian, and geocentric latitude from the earth's equator.
$\hat{r}, \hat{\theta}, \hat{\phi}$	Unit vectors for the spherical coordinate system: r, θ, ϕ.
λ_0 or λ_o	The initial geographic longitude of the near synchronous satellite (i.e., the geographic longitude at injection).
q	An index for the coefficients of the drift equations (57). Also, a characteristic solution of the uncoupled equations of motion (36) (i.e., see equation 59).
ϵ	A small parameter.
i	In addition to a perturbation constant, used as an index to the coefficients of the drift equations (57).
$\mathbf{F}_p$	A perturbation force on the near synchronous satellite.
$\mathbf{P}_a$	The angular momentum of the near synchronous satellite.
$\delta\mathbf{P}_a$	A small change in the angular momentum of the near synchronous satellite due to F_p acting for a short time.
g_s	The acceleration of earth gravity at the near synchronous radius r_s:

$$g_s \doteq 32.15(R_o/r_s)^2 \doteq 32.15 \times 0.02288 = 0.7355 \text{ Ft./Sec.}^2$$

INTRODUCTION AND CHAPTER SUMMARIES

Much analytic work has been done recently on the motions of a near synchronous satellite of an earth with a triaxial (2nd order) gravity field.[1-3] The oscillatory movement of such a satellite about the minor axis of the earth's elliptical equator has been sufficiently well described. This investigation into higher order gravity effects on 24-hour satellites was prompted by a recent refinement of knowledge about these higher order anomalies.[4] The perturbation forces arising from the higher order tesseral harmonics of the earth's gravity field are small, but they are in resonance on a 24-hour satellite. The major conclusion of this investigation is that, while no absolutely stationary geographic points exist for 24-hour satellites in an earth field to 4th order, the overall features of the regime of motion in the dominant triaxial field (see ref. 2) still hold.

In part 1, the coupled linear system of initial perturbational motion for an equatorial 24-hour satellite in an earth field to 4th order (with a circular reference orbit) is derived. The method of derivation follows that in references 1 and 2. The three transcendental equations for absolute stability in the 4th order field are presented and the two dominating equations are solved to determine four sets of injection radii and geographic longitudes at which 24-hour satellites may be placed and maintained with very low drift rates. Approximate integrals of the linear system are found to describe the drift at all longitudes (equations 57).

In part 2, the inherent stability of the motion about the 4 low drift points determined in part 1, is investigated; following the method of reference 2.

In part 3, two examples of the use of the ap-

proximate initial drift integrals (57) are given. A case of perfect injection to give near maximum drift rates, is solved and compared with the numerically integrated drift for the same injection conditions.

Appendix A contains formulas evaluating the coefficients of the drift integrals (57) for a wide range of injection conditions.

Appendix C gives a chart of the variations of the longitude and latitude perturbation forces around the equator at near synchronous altitudes.

Appendix D presents the equations and procedure for determining the injection radius and longitude into a 24-hour orbit with minimal initial perturbations, valid for any 4th order earth field with small anomalies.

DERIVATION OF THE INTEGRALS OF PERTURBATIONAL MOTION FOR A 24-HOUR EARTH SATELLITE

The force field **F** of the earth on a mass point m at earth centered r, ϕ, θ in inertial space (see fig. 1) can be written as;

$$\vec{F} = m\left\{\hat{r}\,\frac{\partial V_E}{\partial r} + \frac{\hat{\phi}}{r}\,\frac{\partial V_E}{\partial \phi} + \frac{\hat{\theta}}{r\cos\phi}\,\frac{\partial V_E}{\partial \theta}\right\}, \tag{1}$$

where

$$V_E = \frac{\mu_E}{r}\sum_{n=2}^{4}\sum_{m=0}^{n}\left[1-\left(\frac{R_0}{r}\right)^n P_n^m(\sin\phi)\,J_{nm}\cos m\left(\theta-\theta_{nm}\right)\right] \tag{2}$$

See appendix B, figure 1 below, and the nomenclature for the significance of the symbols and indices in (1) and (2).

The XY plane is the earth's equatorial plane. θ_{nm} is the inertial longitude of the principal axis of symmetry of the earth's mass distribution accounted for by the nm harmonic of the geopotential V_E.

The acceleration of m in inertial space r, ϕ, θ is;

$$\vec{a} = \hat{r}\left(\ddot{r} - r\dot{\theta}^2\cos^2\phi - r\dot{\phi}^2\right) + \hat{\phi}\left[\left(\frac{1}{r}\right)\left(\frac{d}{dt}\left[r^2\dot{\phi}\right]\right) + r\dot{\theta}^2\cos\phi\,\sin\phi\right]$$

$$+ \hat{\theta}\left[\frac{1}{r\cos\phi}\,\frac{d}{dt}\left(r^2\dot{\theta}\cos^2\phi\right)\right]. \tag{3}$$

Writing $\mathbf{F}=m\mathbf{a}$, implies the following three scalar equations in the r, ϕ, θ components of force and

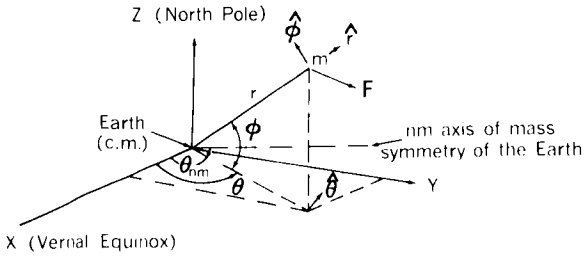


FIG. 1. Coordinate system referencing the motion of a 24-hr earth satellite.

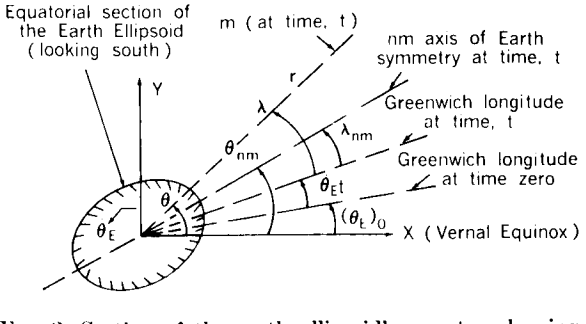


FIG. 2. Section of the earth ellipsoid's equator showing the relationship of the various longitude references used in this report.

acceleration:

$$\ddot{r} - r\dot{\theta}^2 \cos^2\phi - r\dot{\phi}^2 = G_r\left(r, \phi, \theta - \theta_{nm}\right), \tag{4}$$

$$\frac{1}{r}\frac{d}{dt}\left(r^2\dot{\phi}\right) + r\dot{\theta}^2 \cos\phi \sin\phi = G_\phi\left(r, \phi, \theta - \theta_{nm}\right), \tag{5}$$

$$\frac{1}{r\cos\phi}\frac{d}{dt}\left(r^2\dot{\theta}\cos^2\phi\right) = G_\theta\left(r, \phi, \theta - \theta_{nm}\right), \tag{6}$$

The G's are gravitational force components per unit mass (see app. A).

Consider the XY equatorial plane of the earth, with the earth's equator reflecting the mass distribution due to the nm harmonic of the geopotential (fig. 2). λ is the geographic longitude of m. B_{nm} is the geographic longitude of the principal nm axis of earth symmetry. It is clear from figure 2 that $\lambda - B_{nm} = \theta - \theta_{nm}$. The potential in appendix A is thus consistent with that in reference 4.

The reference orbit for the synchronous satellite is a circle in the equatorial plane of radius r_s, traversed at the earth's rotation rate. Therefore, we assume a perturbation solution to (4), (5), and (6) of the following form:

θ = [Initial Inertial Longitude] + [(Earth Rate) (Time)] + [Geographic Longitude Perturbation]

$$= \left[\left(\theta_E\right)_0 + \lambda_0\right] + \left[\dot{\theta}_E t\right] + \left[\Delta\lambda\right]; \tag{7}$$

$$\phi = \Delta\phi; \tag{8}$$

$$r = r_s + \Delta r, \tag{9}$$

$\dot{\theta}_E$ is the time constant earth rotation rate. λ_0 is a constant equal to the initial geographic longitude position of m. $\Delta\lambda$ is the change in geographic longitude with time. Small r_s is a time constant initial radius (the "synchronous radius") of m, to be determined later. Δr is the change in radius of m with time. $(\theta_E)_0$ is the initial inertial longitude of Greenwich. The perturbation (7) replaces $\theta - \theta_{nm}$ in the force field of appendix A, by $(\theta_E)_0 + \lambda_0 + \dot{\theta}_E t + \Delta\lambda + \theta_{nm}$. But $\theta_{nm} = (\theta_E)_0 + \dot{\theta}_E t + B_{nm}$, from figure 2. Therefore, the force field longitude arguments are, for the perturbation solutions;

$$\lambda_0 - \lambda_{nm} + \Delta\lambda.$$

$$\lambda_0 - \lambda_{nm} = \gamma_{nm} \tag{10}$$

define new harmonic constants referred to the initial geographic longitude of m. Thus (7), (8), and (9) are solutions to (4), (5), and (6) if;

$$\Delta\ddot{r} - \left(r_s + \Delta r\right)\left(\dot{\theta}_E + \Delta\dot{\lambda}\right)^2 \cos^2\phi - \left(\dot{r}_s + \Delta\dot{r}\right)\left(\Delta\dot{\phi}\right)^2 = G_r\left(r_s + \Delta r, \Delta\phi, \Delta\lambda + \gamma_{nm}\right), \tag{11}$$

$$\frac{1}{\left(r_s + \Delta r\right)}\frac{d}{dt}\left(r_s + \Delta r\right)^2\left(\Delta\dot{\phi}\right) + \left(r_s + \Delta r\right)\left(\dot{\theta}_E + \Delta\dot{\lambda}\right)^2 \cos\Delta\phi \sin\Delta\phi$$

$$= G_\phi\left(r_s + \Delta r, \Delta\phi, \Delta\lambda + \gamma_{nm}\right), \tag{12}$$

$$\frac{1}{(r_s+\Delta r)\cos\phi}\frac{d}{dt}(r_s+\Delta r)^2\left(\dot{\theta}_E+\Delta\dot{\lambda}\right)\cos^2\Delta\phi = G_\theta\left(r_s+\Delta r, \Delta\phi, \Delta\lambda+\gamma_{nm}\right). \tag{13}$$

Let $w=\dot{\theta}_E$.

Carrying through the indicated differentiations and substitutions, and ignoring all products of perturbations and/or perturbation rates (11)–(13) becomes;

$$\ddot{r}_1 + ar_1 + b\dot{\lambda}_1 + c\lambda_1 + d\phi_1 = e \tag{14}$$

$$\ddot{\phi}_1 + f\phi_1 + gr_1 + h\lambda_1 = i \tag{15}$$

$$\ddot{\lambda}_1 + j\lambda_1 + k\dot{r}_1 + \ell r_1 + m\phi_1 = n\,, \tag{16}$$

The dots in (14)–(16) refer to derivatives with respect to a dimensionless time, $T=tw$. Thus;

$$\frac{d}{dT} = \frac{1}{w}\frac{d}{dt}; \qquad \frac{d^2}{dT^2} = \frac{1}{w^2}\frac{d^2}{dt^2}.$$

The constants in (14)–(16) are:

$$a = -1-\frac{\mu_E}{w^2r_s^3}\left\{2+6J_{20}\left(\frac{R_0}{r_s}\right)^2 - 36J_{22}\left(\frac{R_0}{r_s}\right)^2\cos 2\gamma_{22} + 30\left(\frac{R_0}{r_s}\right)^3 J_{31}\cos\gamma_{31}\right.$$

$$- 300J_{33}\left(\frac{R_0}{r_s}\right)^2\cos 3\gamma_{33} - 11.25J_{40}\left(\frac{R_0}{r_s}\right)^4 + 225J_{42}\left(\frac{R_0}{r_s}\right)^4\cos 2\gamma_{42}$$

$$\left. - 3150J_{44}\left(\frac{R_0}{r_s}\right)^4\cos 4\gamma_{44}\right\}; \tag{17}$$

$$b = -2; \tag{18}$$

$$c = \frac{\mu_E}{w^2r_s^3}\left\{18J_{22}\left(\frac{R_0}{r_s}\right)^2\sin 2\gamma_{22} - 6J_{31}\left(\frac{R_0}{r_s}\right)^3\sin\gamma_{31} + 180J_{33}\left(\frac{R_0}{r_s}\right)^3\sin 3\gamma_{33}\right.$$

$$\left. - 75J_{42}\left(\frac{R_0}{r_s}\right)^4\sin 2\gamma_{42} + 2100J_{44}\left(\frac{R_0}{r_s}\right)^4\sin 4\gamma_{44}\right\}; \tag{19}$$

$$d = \frac{\mu_E}{w^2 r_s^3}\left\{6J_{30}\left(\frac{R_0}{r_s}\right)^3 - 60J_{32}\left(\frac{R_0}{r_s}\right)^3 \cos 2\gamma_{32} + \frac{75}{2} J_{41}\left(\frac{R_0}{r_s}\right)^4 \cos \gamma_{41}\right.$$

$$\left. - 525J_{43}\left(\frac{R_0}{r_s}\right)^4 \cos 3\gamma_{43}\right\}; \tag{20}$$

$$e = 1 + \frac{\mu_E}{w^2 r_s^3}\left\{-1 - \frac{3}{2} J_{20}\left(\frac{R_0}{r_s}\right)^2 + 9J_{22}\left(\frac{R_0}{r_s}\right)^2 \cos 2\gamma_{22} - 6J_{31}\left(\frac{R_0}{r_s}\right)^3 \cos \gamma_{31}\right.$$

$$+ 60J_{33}\left(\frac{R_0}{r_s}\right)^3 \cos 3\gamma_{33} + \frac{15}{8} J_{40}\left(\frac{R_0}{r_s}\right)^4 - \frac{75}{2} J_{42}\left(\frac{R_0}{r_s}\right)^4 \cos 2\gamma_{42}$$

$$\left. + 525J_{44}\left(\frac{R_0}{r_s}\right)^4 \cos 4\gamma_{44}\right\}; \tag{21}$$

$$f = 1 - \frac{\mu_E}{w^2 r_s^3}\left\{-3J_{20}\left(\frac{R_0}{r_s}\right)^2 + 6J_{22}\left(\frac{R_0}{r_s}\right)^2 \cos 2\gamma_{22} - \frac{32}{2} J_{31}\left(\frac{R_0}{r_s}\right)^3 \cos \gamma_{31}\right.$$

$$+ 45J_{33}\left(\frac{R_0}{r_s}\right)^3 \cos 3\gamma_{33} + \frac{15}{2} J_{40}\left(\frac{R_0}{r_s}\right)^4 - 120J_{42}\left(\frac{R_0}{r_s}\right)^4 \cos 2\gamma_{42}$$

$$\left. + 420J_{44}\left(\frac{R_0}{r_s}\right)^4 \cos 4\gamma_{44}\right\}; \tag{22}$$

$$g = -\frac{\mu_E}{w^2 r_s^3}\left\{-9J_{30}\left(\frac{R_0}{r_s}\right)^3 + 90J_{32}\left(\frac{R_0}{r_s}\right)^3 \cos 2\gamma_{32} - \frac{105}{2} J_{41}\left(\frac{R_0}{r_s}\right)^4 \cos \gamma_{41}\right.$$

$$\left. + 735J_{43}\left(\frac{R_0}{r_s}\right)^4 \cos 3\gamma_{43}\right\}; \tag{23}$$

$$h = -\frac{\mu_E}{w^2 r_s^3}\left\{30J_{32}\left(\frac{R_0}{r_s}\right)^3 \sin 2\gamma_{32} - \frac{15}{2} J_{41}\left(\frac{R_0}{r_s}\right)^4 \sin \gamma_{41}\right.$$

$$\left. + 315J_{43}\left(\frac{R_0}{r_s}\right)^4 \sin 3\gamma_{43}\right\}; \tag{24}$$

$$i = \frac{\mu_E}{w^2 r_s^3}\left\{\frac{15}{2} J_{30}\left(\frac{R_0}{r_s}\right)^3 - 15 J_{32}\left(\frac{R_0}{r_s}\right)^3 \cos 2\gamma_{32} + \frac{15}{2} J_{41}\left(\frac{R_0}{r_s}\right)^4 \cos \gamma_{41} - 105 J_{43}\left(\frac{R_0}{r_s}\right)^4 \cos 3\gamma_{43}\right\}; \tag{25}$$

$$j = -\frac{\mu_E}{w^2 r_s^3}\left\{12 J_{22}\left(\frac{R_0}{r_s}\right)^2 \cos 2\gamma_{22} - \frac{3}{2} J_{31}\left(\frac{R_0}{r_s}\right)^3 \cos \gamma_{31} + 135 J_{33}\left(\frac{R_0}{r_s}\right)^3 \cos 3\gamma_{33} - 30 J_{42}\left(\frac{R_0}{r_s}\right)^4 \cos 2\gamma_{42} + 1680 J_{44}\left(\frac{R_0}{r_s}\right)^4 \cos 4\gamma_{44}\right\}; \tag{26}$$

$$k = 2. \tag{27}$$

$$l = -\frac{\mu_E}{w^2 r_s^3}\left\{-30 J_{22}\left(\frac{R_0}{r_s}\right)^2 \sin 2\gamma_{22} + 9 J_{31}\left(\frac{R_0}{r_s}\right)^3 \sin \gamma_{31} - 270 J_{33}\left(\frac{R_0}{r_s}\right)^3 \sin 3\gamma_{33} + 105 J_{42}\left(\frac{R_0}{r_s}\right)^4 \sin 2\gamma_{42} - 2940 J_{44}\left(\frac{R_0}{r_s}\right)^4 \sin 4\gamma_{44}\right\}; \tag{28}$$

$$m = -\frac{\mu_E}{w^2 r_s^3}\left\{+30 J_{32}\left(\frac{R_0}{r_s}\right)^3 \sin 2\gamma_{32} - \frac{15}{2} J_{41}\left(\frac{R_0}{r_s}\right)^4 \sin \gamma_{41} + 315 J_{43}\left(\frac{R_0}{r_s}\right)^4 \sin 3\gamma_{43}\right\}; \tag{29}$$

$$n = \frac{\mu_E}{w^2 r_s^3}\left\{6 J_{22}\left(\frac{R_0}{r_s}\right)^2 \sin 2\gamma_{22} - \frac{15}{2} J_{31}\left(\frac{R_0}{r_s}\right)^3 \sin \gamma_{31} + 45 J_{33}\left(\frac{R_0}{r_s}\right)^3 \sin 3\gamma_{33} - 15 J_{42}\left(\frac{R_0}{r_s}\right)^4 \sin 2\gamma_{42} + 420 J_{44}\left(\frac{R_0}{r_s}\right)^4 \sin 4\gamma_{44}\right\}. \tag{30}$$

Writing (14)–(16) in operator notation; ($s^1=d/dT$, $s^2=d^2/dT^2$, $s=s^1$, $s^0=1$, etc.)

$$\left(s^2+a\right) r_1 + (bs+c)\,\lambda_1 + (d)\,\phi_1 = e\,, \tag{31}$$

$$(g)\, r_1 + (h)\,\lambda_1 + \left(s^2+f\right)\phi_1 = i\,, \tag{32}$$

$$(ks+l)\, r_1 + \left(s^2+j\right)\lambda_1 + (m)\,\phi_1 = n\,. \tag{33}$$

In (14)–(16) and (31)–(33); $r_1=\Delta r/r_s$, $\lambda_1=\Delta\lambda$, $\phi_1=\Delta\phi$: all dimensionless variables.
Solving (31)–(33) by Cramer's rule:

$$r_1 = \frac{\begin{vmatrix} e & bs+c & d \\ i & h & s^2+f \\ n & s^2+j & m \end{vmatrix}}{\begin{vmatrix} s^2+a & bs+c & d \\ g & h & s^2+f \\ ks+l & s^2+j & m \end{vmatrix}} = \Big\{ e\left[hm - s^4 - jf - s^2(f+j)\right] - i\left[m(bs+c) \right.$$

$$\cdots - d\left(s^2+j\right)\Big] + n\left(bs^3+cs^2+bfs+cf-h\right)\Big\}\Big/\Big\{\left[s^2+a\right]\left[mh-s^4\right.$$

$$\cdots - s^2(j+f) - jf\Big] - g\left[m(bs+c) - d\left(s^2+j\right)\right] + (ks+l)\left(bs^3+cs^2\right.$$

$$\left.\cdots + bfs + cf - hd\right)\Big\}. \tag{34}$$

Or

$$-s^6 r_1 + (A)\,s^4 r_1 + (B)\,s^3 r_1 + (C)\,s^2 r_1 + (D)\,s^1 r_1 + (E)\,s^0 r_1 = ehm + ejf - imc \cdots$$

$$\cdots + idj + ncf - nhd. \tag{35}$$

This result follows from the evaluation of the determinants because $s(a, b, c, \cdots)=0$, since a, b, c $\cdots$ are all constants. Summarizing, the three uncoupled linear equations of 6th order are:

$$\left[-s^6 + (A)\,s^4 + (B)\,s^3 + (C)\,s^2 + (D)\,s^1 + (E)\,s^0\right] r_1 = A_1\,,$$

$$\left[-s^6 + (A)\,s^4 + (B)\,s^3 + (C)\,s^2 + (D)\,s^1 + (E)\,s^0\right] \lambda_1 = A_2\,, \tag{36}$$

$$\left[-s^6 + (A)\,s^4 + (B)\,s^3 + (C)\,s^2 + (D)\,s^1 + (E)\,s^0\right] \phi_1 = A_3\,.$$

The constants are;

$$A = -a + kb - j - f , \tag{37}$$

$$B = kc + lb , \tag{38}$$

$$C = mh - aj - af - jf + gd + lc + bfk , \tag{39}$$

$$D = -gmb + kcf + lbf - khd , \tag{40}$$

$$E = amh - ajf - gmc + dgj + lcf - lhd , \tag{41}$$

$$A_1 = ehm + ejf - imc + idj + ncf - nhd , \tag{42}$$

$$A_2 = aim - anf - gem + gnd + lef - lid , \tag{43}$$

$$A_3 = ahn - aij - gcn + gej + lci - leh . \tag{44}$$

Examination of (31)–(33) shows that for zero natural initial conditions; ($r_1=\dot{r}_1=\phi_1=\dot{\phi}_1=\lambda_1=\dot{\lambda}_1=0$, at $T=0$) $\ddot{r}_1(T=0)=e$, $\ddot{\lambda}_1(T=0)=n$ and $\dot{\phi}_1(T=0)=i$. The necessary and sufficient conditions for the perturbations to be zero for all time are, then, for the mass m to be placed, with zero initial conditions, into an orbit for which:

$$e=n=i=0. \tag{45}$$

That this is so may be shown by successive differentiation of (31)–(33) for the higher derivatives. They will all be zero providing only that the initial perturbation, perturbation rates and perturbation accelerations are zero. But (45) are essentially three transcendental equations in the two unknowns λ_0 and r_s (the initial longitude and radius of the satellite). Therefore, there will be, in general, no simultaneous solution except by coincidence of the constants of those equations. However, from what is known at present about the earth's gravitational field (see ref. 4 and app. C), the perturbation forces due to the latitude antisymmetry of the field (included in the i constant) are small compared to those in the radial and longitudinal directions at near synchronous altitudes. The latitude perturbations, then, may be neglected in considering the conditions for a near zero solution to (31)–(33). A plot of these perturbation forces with λ_0 at an r_s determined from $e=0$, is found in appendix C. It is postulated then, and proved later, that essentially stable regions of the gravity field in geographic-geocentric coordinates exist in the neighborhood of one or more points on the equator for which;

$$e=n=0. \tag{46}$$

In the development which follows, the earth gravity field of ref. *4* is used together with an earth rotation rate from reference *5*.
These earth constants are:

$$
\left.
\begin{aligned}
w &= .7292115 \times 10^{-4} \text{ rad./sec.} \qquad \text{(Reference 3)} \\
\mu_E &= 3.986032 \times 10^{20} \text{ cm.}^3/\text{sec.}^2 \\
R_0 &= 6.378165 \times 10^{8} \text{ cm.} \\
J_{20} &= 1082.48 \times 10^{-6} \\
J_{30} &= -2.56 \times 10^{-6} \\
J_{40} &= -1.84 \times 10^{-6} \\
J_{21} &= 0. \\
J_{22} &= -1.2 \times 10^{-6} \\
J_{31} &= -1.9 \times 10^{-6} \\
J_{32} &= -.14 \times 10^{-6} \\
J_{33} &= -.10 \times 10^{-6} \\
J_{41} &= -.52 \times 10^{-6} \\
J_{42} &= -.062 \times 10^{-6} \\
J_{43} &= -.035 \times 10^{-6} \\
J_{44} &= -.031 \times 10^{-6} \\
\lambda_{22} &= -26.4 \text{ degrees} \\
\lambda_{31} &= 4.6 \text{ degrees} \\
\lambda_{32} &= -16.8 \text{ degrees} \\
\lambda_{33} &= 42.6 \text{ degrees} \\
\lambda_{41} &= 237.5 \text{ degrees} \\
\lambda_{42} &= 65.2 \text{ degrees} \\
\lambda_{43} &= 0.5 \text{ degrees} \\
\lambda_{44} &= 14.9 \text{ degrees}
\end{aligned}
\right\} \qquad (47)
$$

The two transcendental equations in λ_0 and r_s arising from (46) are:

$$\frac{w^2 r_s^3}{\mu_E} = 1 + \frac{3}{2} J_{20}\left(\frac{R_0}{r_s}\right)^2 + \cos\lambda_0 \left\{-6J_{31}\left(\frac{R_0}{r_s}\right)^3 \cos\lambda_{31}\right\} + \cos 2\lambda_0 \left\{-9J_{22}\left(\frac{R_0}{r_s}\right)^2 \cos 2\lambda_{22} + \frac{75}{2} J_{42}\left(\frac{R_0}{r_s}\right)^4\right\} + \cos 3\lambda_0 \left\{-60J_{33}\left(\frac{R_0}{r_s}\right)^3 \cos 3\lambda_{33}\right\} + \cos 4\lambda_0 \cdot \left\{-525J_{44}\left(\frac{R_0}{r_s}\right)^4 \cos 4\lambda_{44}\right\} - \frac{15}{8} J_{40}\left(\frac{R_0}{r_s}\right)^4 + \sin\lambda_0 \left\{-6J_{31}\left(\frac{R_0}{r_s}\right)^3 \sin\lambda_{31}\right\} + \sin 2\lambda_0 \left\{-9J_{22}\left(\frac{R_0}{r_s}\right)^2 \sin 2\lambda_{22} + \frac{75}{2} J_{42}\left(\frac{R_0}{r_s}\right)^4 \sin 2\lambda_{42}\right\} + \sin 3\lambda_0 \left\{-60J_{33}\left(\frac{R_0}{r_s}\right)^3 \sin 3\lambda_{33}\right\} + \sin 4\lambda_0 \left\{-525J_{44}\left(\frac{R_0}{r_s}\right)^4 \sin 4\lambda_{44}\right\}; \quad (48)$$

$$0 = \sin\lambda_0 \left[-\frac{3}{2} J_{31}\left(\frac{R_0}{r_s}\right) \cos\lambda_{31}\right] + \sin 2\lambda_0 \left[6J_{22} \cos 2\lambda_{22} - 15J_{42}\left(\frac{R_0}{r_s}\right)^2\right] + \sin 3\lambda_0 \left[45J_{33}\left(\frac{R_0}{r_s}\right) \cos 3\lambda_{33}\right] + \sin 4\lambda_0 \left[420J_{44}\left(\frac{R_0}{r_s}\right)^2 \cos 4\lambda_{44}\right] + \cos\lambda_0 \left[\frac{3}{2} J_{31}\left(\frac{R_0}{r_s}\right) \sin\lambda_{31}\right] + \cos 2\lambda_0 \left[-6J_{22} \sin 2\lambda_{22} + 15J_{42}\left(\frac{R_0}{r_s}\right)^2 \sin 2\lambda_{42}\right] + \cos 3\lambda_0 \left[-45J_{33}\left(\frac{R_0}{r_s}\right) \sin 3\lambda_{33}\right] + \cos 4\lambda_0 \left[-420J_{44}\left(\frac{R_0}{r_s}\right)^2 \sin 4\lambda_{44}\right]. \quad (49)$$

In general, for every finite zero r_s, (49) changes sign a minimum of two times and a maximum of eight times over the equator. The earth constants are such that for $r_s > R_0$, the right-hand side of (48) is very close to 1 for all points on the equator. Essentially then, (48) and (49) decouple. Equation (48) may be solved separately for a near synchronous radius independent of longitude, leaving a small longitude dependent residual. With the near synchronous radius so determined, the zero's of (49) establish (to high accuracy) a minimum of two and a maximum of 8 potentially stable points on the equator. The longitude dependent residual of (48) may then be solved to establish (to high accuracy) the radius to each potentially stable point. The perturbation con-

stants of the earth's gravity field are so much less than 1., that the iteration need not be carried further than that outlined above (see app. D). The results of this iteration of (48) and (49) with the earth constants of (47) are: The spherical earth "synchronous radius" is;

r_s(Spherical)

$$=138333942.5 \text{ Ft. } (26199.6103 \text{ St. Miles}). \quad (50)$$

The "oblate earth" (including J_{20} and J_{40} potential terms) "synchronous radius" is;

r_s(Oblate)

$$=138335648.5 \text{ Ft. } (26199.9334 \text{ St. Miles}). \quad (51)$$

With the "oblate earth" synchronous radius above, (49) becomes:

$$0=.430\sin\lambda_0-4.37\sin2\lambda_0+0.4175\sin3\lambda_0 -.151\sin4\lambda_0-0.0345\cos\lambda_0-5.76\cos2\lambda_0 +0.538\cos3\lambda_0+0.257\cos4\lambda_0. \quad (52)$$

The zero's of (52) which are the potentially stable points around the equator, are at;

$$\lambda_0=64.2°,\ 155.8°,\ 242.9° \text{ and } 331.3°. \quad (53)$$

It is interesting to compare these potentially stable longitudes with those which would be present if all the tesseral and sectorial harmonics except J_{22} are ignored. In the simpler field (the so-called "triaxial" gravity field), the zero's of (48) with the oblate earth synchronous radius of (51) are at;

$$\lambda_0=63.6°,\ 153.6°,\ 243.6° \text{ and } 333.6°. \quad (54)$$

In no case (with a full earth potential) do the "stable" points differ by more than $2.3\dot{r}$ from those which exist in the simpler "triaxial" field.

The potentially stable radii to the longitudes of (53) are:

$r_s(\lambda_0=64.2°)=138335637.5$ Ft.

$$\times(26199.9313 \text{ St. Miles}) \quad (55\text{–A})$$

$r_s(\lambda_0=155.8°)=138335660.2$ Ft.

$$\times(26199.9356 \text{ St. Miles}) \quad (55\text{–B})$$

$r_s(\lambda_0=242.9°)=138335635.2$ Ft.

$$\times(26199.9309 \text{ St. Miles}) \quad (55\text{–C})$$

$r_s(\lambda_0=331.3°)=138335660.3$ Ft.

$$\times(26199.9356 \text{ St. Miles}). \quad (55\text{–D})$$

The uncoupled linearized equations (36) may be simplified and integrated directly by ignoring all terms in the differential coefficients A, B, C, D and E which are much less than 1. The driving terms A_1, A_2, A_3 must retain at least one order of smallness less than 1 so that the resulting solution is sufficiently sensitive to drift acceleration. The initial radius may be chosen for convenience as the mean of those in (55). But to insure the longest possible validity for the resulting solution, it is probably best to solve $e=0$ for r_s at the λ_0 from which the perturbation is desired. In any case, $r_s=138335647.7\pm12.5$ ft for near zero solutions to (36) with zero initial conditions. The simplified uncoupled equations then become;

$$\begin{aligned} (s^6+2s^4+s^2)\,r_1 &= -A_1 \\ (s^6+2s^4+s^2)\,\lambda_1 &= -A_2 \\ (s^6+2s^4+s^2)\,\phi_1 &= -A_3\,. \end{aligned} \quad (56)$$

It may be verified that the complete solution to (56) is:

$$\Delta_q = C_{1q}+C_{2q}T+(C_{3q}+C_{4q}T)\sin T+(C_{5q}+C_{6q}T)\cos T-\frac{A_qT^2}{2}\,, \quad (57)$$

where:

$$\Delta_q = r_1 \text{ when } q = 1,$$

$$\Delta_q = \lambda_1 \text{ when } q = 2,$$

$$\Delta_q = \phi_1 \text{ when } q = 3.$$

In any dynamics problem utilizing the perturbation solutions (57), 18 conditions on the perturbations must be specified. The natural conditions of initial position and velocity are the one's most often encountered in practice, generating the motion away from the reference position. For this problem, the zero'th through fifth derivative of (57) for each coordinate perturbation must be evaluated at $T=0$. The zero'th and first derivatives of the perturbations at $T=0$ (s_o^0 and s_o^1) are arbitrary, but must be small for this linear solution to be valid for a reasonable length of time. The second through the fifth derivatives of the perturbations are evaluated successively from (14)–(16) once the reference radius (r_s) and longitude (λ_0) are specified. The reference latitude is zero. The coefficients $C_{iq}(i=1, 2, \cdots 6; q=1, 2, 3)$ are then solved from the resulting system of equations and appear in appendix A.

INVESTIGATION OF THE STABILITY OF MOTIONS IN THE VICINITY OF POINTS AT WHICH THE LONGITUDINAL AND RADIAL PERTURBATION FORCES ARE ZERO

If the longitudinal and radial perturbation forces are zero, it may be verified that $e=n=0$ (see app. B and C) and the linearized equations (31)–(33) become;

$$\left.\begin{aligned} (s^2+a)r_1 + (bs+c)\lambda_1 + (d)\phi_1 &= 0; \\ (g)r_1 + (h)\lambda_1 + (s^2+f)\phi_1 &= i; \\ (ks+l)r_1 + (s^2+j)\lambda_1 + (m)\phi_1 &= 0. \end{aligned}\right\} \tag{58}$$

The characteristic equation of 58 is; (evaluating the differential constants of (36) for $e=n=0$)

$$q^6(-1) + q^4(-a+kb-j-f) + q^3(kc+lb) + q^2(mh-aj-af-jf+gd+lc+bfk)$$

$$+ q(-gmb+kcf+lbf-khd) + amh + dgj + lcf - lhd = 0. \tag{59}$$

The particular solutions to (58) will be constants:

$$\begin{aligned} r_1 \text{ (particular)} &= E/A_1 \\ \lambda_1 \text{ (particular)} &= E/A_2 \\ \phi_1 \text{ (particular)} &= E/A_3. \end{aligned} \tag{60}$$

To simplify the calculation of the roots of the characteristic equation (59) without losing anything essential in the characteristics of the motion, we will ignore the lower order perturbation constants and/or products of perturbation constants in each coefficient of (59). With this simplification and with the large constants evaluated, (59) goes to:

$$q^6(-1) + q^4(-2) + q^3(2c-2l) + q^2(-1) + q(2c-2l) + 3j = 0. \tag{61}$$

It is noted that the characteristic values of the motion from (61) are independent of the latitude perturbations of the earth's gravity field. Consider $q_{1,2}$ as solutions to (61), where $q_{1,2} \ll 1$. For these solu-

tions, ignore orders of q smaller than q^2. Equation 61 in $q_{1,2}$ then becomes;

$$q^2 + 2(l-c)q - 3j = 0 . \tag{62}$$

The solutions to (62) are:

$$q_{1,2} = -(l-c) \pm \left[(l-c)^2 + 3j\right]^{1/2} . \tag{63}$$

Since $|l-c|$, $|3j| \ll 1$; the approximation to (62) for these roots is valid.

There are also solutions to (61); $|q| \doteq (-1)^{1/2}$. Therefore, let;

$$q_{3,4} = +(-1)^{1/2} + \epsilon_{3,4} , \tag{64}$$

where $|\epsilon_{3,4}| \ll 1$. With (64) in (61), (61) becomes;

$$\left[(-1)^{1/2} + \epsilon_{3,4}\right]^6 + 2\left[(-1)^{1/2} + \epsilon_{3,4}\right]^4 + 2(l-c)\left[(-1)^{1/2} + \epsilon_{3,4}\right]^3$$

$$+ \left[(-1)^{1/2} + \epsilon_{3,4}\right]^2 + \left[(-1)^{1/2} + \epsilon_{3,4}\right]\left[2(l-c)\right] - 3j = 0 . \tag{65}$$

Since $|\epsilon_{3,4}| \ll 1$, the expansion of (65) ignoring terms in ϵ^3 and higher, is:

$$6\epsilon\left[(-1)^{1/2}\right]^5 + 8\epsilon\left[(-1)^{1/2}\right]^3 + 6(l-c)\epsilon\left[(-1)^{1/2}\right]^2 + 2\epsilon(-1)^{1/2} - 2(l-c)(-1)^{1/2}$$

$$+ 2\epsilon(l-c)(-1)^{1/2} + 2(l-c)(-1)^{1/2} + 15\epsilon^2 - 12\epsilon^2 + 6(l-c)\epsilon^2(-1)^{1/2}$$

$$+ \epsilon^2 = 0 .$$

This last equation reduces to:

$$\epsilon^2\left[15 - 12 + 6(l-c)(-1)^{1/2} + 1\right] + \epsilon\left[6(-1)^{1/2} - 8(-1)^{1/2} - 6(l-c) + 2(-1)^{1/2}\right.$$

$$\left. + 2(l-c)(-1)^{1/2}\right] = 0 .$$

Thus, the approximate roots to (65) are:

$$\epsilon_3 = 0 , \qquad \epsilon_4 = \frac{-2(l-c)\left[-3 + (-1)^{1/2}\right]}{4} . \tag{66}$$

Similarly, letting:

$$q_{5,6} = -(-1)^{1/2} + \epsilon_{5,6} . \tag{67}$$

The expansion of (61), with the same order of approximation as for $\epsilon_{3,4}$, reduces to:

$$6\epsilon\left[-(-1)^{1/2}\right]^5 + 8\epsilon\left[-(-1)^{1/2}\right]^3 + 6(l-c)\epsilon\left[-(-1)^{1/2}\right]^2 + 2\epsilon\left[-(-1)^{1/2}\right] + 2(l-c)(-1)^{1/2}$$

$$- 2\epsilon(l-c)(-1)^{1/2} - 2(l-c)(-1)^{1/2} + 15\epsilon^2 - 12\epsilon^2 - 6(l-c)\epsilon^2(-1)^{1/2} + \epsilon^2 = 0 .$$

The roots to the above equation are:

$$\epsilon_5 = 0 , \qquad \epsilon_6 \pm \frac{-2(l-c)\left[-(-1)^{1/4}-3\right]}{4} . \tag{68}$$

Thus (66) and (68) in (64) and (67) determine the other four roots to (61) as approximately:

$$q_{3,5} = \pm(-1)^{1/4} , \qquad q_{4,6} = \pm(-1)^{1/4}\left[1-\tfrac{1}{2}(l-c)\right]+\tfrac{3}{2}(l-c) . \tag{69}$$

Stability in the Triaxial Geopotential Field

In this case, $l=c=0$ at the points of zero longitudinal and radial perturbation forces. The four characteristic solutions are:

$$q_{1,2}\ (\text{triaxial}) \pm \pm(3j)^{1/4} , \tag{70-A}$$

$$q_{3,4}\ (\text{triaxial}) \pm \pm(-1)^{1/4} . \tag{70-B}$$

While it is true that there are two more independent complementary solutions to (58) that may be found for the triaxial case, their constants will be found to be zero when a natural set of initial conditions are specified. For the triaxial case, $g=h=d=i=m=0$. Therefore (58) uncouples in the latitude variation and there are only four characteristic solutions to the set in the longitude and radial perturbations. It may be shown that (70–A) and (70–B) are just these four. It is interesting that the redundant $q_{5,6}=\pm(-1)^{1/2}$ are just the two characteristic solutions to the latitude variation in the triaxial case.

On or Near the Major Axis in the Triaxial Field

$3j(\lambda_0=153.6°,\ -26.4°)=+0.99\times10^{-6}$ and the motion thus has a slowly divergent component.

On or Near the Minor Axis in the Triaxial Field

$3j(\lambda_0=63.6°,\ -116.4°)=-0.99°\times10^{-6}$. The motion consists of two nondamped self-limiting oscillations. One has a period in the neighborhood of 1 day (from $q_{3,4}$). The other has a frequency of $(.99)^{1/2}\times10^{-3}=0.995\times10^{-3}$ (dimensionless) = $0.995\times10^{-3}w$ (dimensions of time^{-1}). This long period oscillation has a period in the neighborhood of

$$\frac{2\pi(\text{rad./cycle})}{.995\times10^{-3}\times2\pi(\text{rad./day})} = 1005 \text{ days} = 2.76 \text{ years} .$$

Stability in the Geopotential Field Through 4th Order

For motion in the vicinity of the 4 points of zero longitudinal and radial perturbation forces:

1. There will be four characteristic solutions giving damped oscillations with periods near 1 day ($q_{3,4,5,6}$).

2. There will be two characteristic solutions giving either a weakly negatively damped oscillation with a period near 1000 days, or two exponentials, one of which is slowly divergent in character.

For Motion in the Vicinity of the Zero Perturbations at $\lambda_0=64.2°$: $l-c=-0.36\times10^{-9}$, $3j=-1.026\times10^{-6}$. Perfect injection in the vicinity of this quasi-stable point is thus followed by two self-limiting oscillations with periods in the vicinity of 1 day which are weakly positively damped ($q_{3,4,5,6}$). There is also a self-limiting oscillation with a long period of $1/(1.026)^{1/2}\times10^{-3}=987$ *Days* $=2.71$ *Years*, which is weakly negatively damped ($q_{1,2}$).

For Motion in the Vicinity of the Zero Perturbations at $\lambda_0=155.8°$: $3j$ is controlled by the J_{22} term, is greater than zero, and $\pm(3j)^{1/2}$ is of the order of $\pm10^{-3}$. $|l-c|$ is of the order of 10^{-9} as

before, so that one of the characteristic solutions will be a slowly divergent exponential.

For Motion in the Vicinity of the Zero Perturbations at $\lambda_0 = 242.9\dot{r}(-117.1\dot{r})$: $l-c = +5.3\times10^{-9}$, $3j = -1.123\times10^{-6}$. Perfect injection in the vicinity of this quasi-stable point is thus followed by two self-limiting coupled oscillations with periods in the vicinity of 1 day which are weakly negatively damped ($q_{3,4,5,6}$). There is also a coupled self-limiting oscillation with a long period near $1/(1.123)^{1/2}\times10^{-3} = 944$ *Days* $= 2.58$ *Years*, which is weakly positively damped ($q_{1,2}$).

For Motion in the Vicinity of the Zero Perturbations at $\lambda_0 = 331.3°(-28.7°)$: $3j$, controlled by the J_{22} term, is greater than zero. $\pm(3j)^{1/2}$ is of the order of $\pm10^{-3}$. $|l-c|$ is of the order of 10^{-9} as before. Therefore, the resultant motion is slowly divergent in character as one of the $q_{1,2}$ solutions will be positive of the order of $+10^{-3}$.

In conclusion, for the geopotential gravity field through 4th order, two regions on the equator have been found within which small initial perturbations in geographic longitude, latitude, and radius are self-limiting in the sense that the resulting motion of an earth satellite in these regions is essentially harmonic in character with very weak damping. These regions are in the neighborhood of 64.2° east of Greenwich and 26,199.9314 statute miles; and 117.1° west of Greenwich and 26,199.9309 statute miles from the center of mass of the earth. They are both within $2\frac{1}{2}°$ of the minor axis of the earth ellipsoid's equator. The damping is of the order of 10^{-9} wt, or of the order of $10^{-8}t$ (days). Thus, initial amplitudes of the damped harmonic perturbations in these two regions suffer a twofold change in magnitude in the order of

$$t = \ln(2)/10^{-8} = 6.93\times10^{7}\ \text{Days} = 1.9\times10^{5}\ \text{Years}.$$

These regions then, can be considered to be essentially stable.

While the regions around 64.2° and 242.9° on the equator have an inherent dynamic stability in the sense shown above, the regions around the zero perturbation points at 155.8° and 331.3° are only very weakly dynamically unstable. This may be shown by calculating the C_{iq} coefficients of the divergent approximate integrals in equation 57. At the $n=e=0$ point where $\lambda_0 = 331.3°$ and $r_s = 138{,}335{,}660.3$ feet, assuming zero initial perturbations and rates, the resultant motion is:

$$r_1 \doteq -1.63\times10^{-16} + (-1.76\times10^{-16}T) + (+2.64\times10^{-16} - .815\times10^{-16}T)\sin T + (1.63\times10^{-16} - .88\times10^{-16}T)\cos T - 21.3\times10^{-24}T^2, \quad (71\text{–A})$$

$$\lambda_1 \doteq -6.15\times10^{-16} + (3.26\times10^{-16}T) + (-4.89\times10^{-16} + 1.76\times10^{-16}T)\sin T + (+6.15\times10^{-16} + 1.63\times10^{-16}T)\cos T + 13.2\times10^{-17}T^2, \quad (71\text{–B})$$

$$\phi_1 \doteq -6.4\times10^{-9} + (-23.7\times10^{-14}T)\sin T + (6.4\times10^{-9}\cos T) + 24.5\times10^{-16}T^2. \quad (71\text{–C})$$

(71–A) predicts a change in injection radius of only -0.0003 foot in 5 years. (71–B) predicts a change in injection geographic longitude of only $+1.00\times10^{-6}$ degrees in 5 years. (71–C) predicts a change in injection latitude of only $+1.85\times10^{-5}$ degrees in 5 years. These are all mean daily drifts. The amplitudes of the daily oscillations are even smaller quantities. The conclusion is that station-keeping requirements for near synchronous satellites placed with low initial rates near the major axis of the earth ellipsoid's equator, will be virtually unaffected by the noncentral character of the earth's gravitational field.

INVESTIGATION OF THE DRIFT OF AN EARTH SATELLITE INJECTED WITH LOW INITIAL RATES AT OR NEAR THE EQUATOR AT ANY LONGITUDE AND AT A RADIUS AT WHICH THE RADIAL PERTURBATION FORCE OF THE EARTH'S GRAVITY FIELD* IS ZERO

The general character of the drift following near perfect injection at radii for which e is equal to or close to zero, may be found by examining the approximate integrals (57) and their coefficients. There is a coupled daily harmonic oscillation in the drift motion due to the eccentricity introduced by the resonant longitudinal perturbation force (when present) principally. This daily harmonic oscillation is also partly due to coupling from the latitude and radial perturbations. The long term drift away from the injection point predicted by (57) is controlled by the C_{2q} and/or A_q terms. It may be seen by examining the order of magnitudes of the coefficients in appendix A, that the long-term drift in radius is controlled by C_{21} which is about $2n$ for perfect injection ($s_o{}^1=s_o{}^0=0$). The long-term drift in longitude is controlled by $-A_2/2$ which is about $-3n/2$ for perfect injection. The sign of n changes four times around the equator. The following table shows the dominant drift effect following perfect injection:

	Sign (n) (dominant drift in radius)	Sign ($-n$) (dominant drift in longitude)
$62.4°<\lambda_o<155.8°$:	+	−
$155.8°<\lambda_o<242.9°$:	−	+ (72)
$242.9°<\lambda_o<331.3°$:	+	−
$331.3°<\lambda_o<62.4°$:	−	+

*See appendix B.

Thus the long-term radial and longitudinal drift following near perfect injection at near synchronous radii ($e\doteq0$) around the equator has the following character: (out of scale)

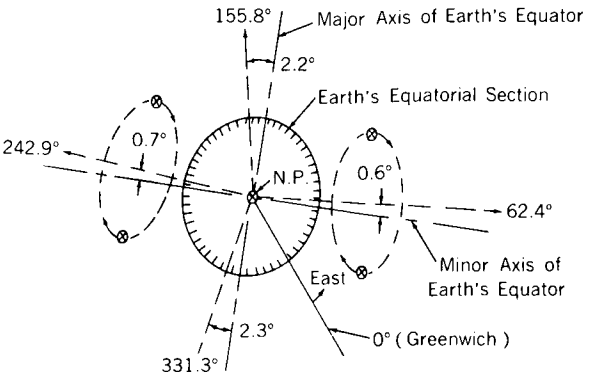


FIG. 3. Long term radial and longitudinal drift following near perfect injection at near synchronous radii around the equator.

The long-term motion in longitude and radius is thus highly suggestive of the coupled long-term circulation about the minor axis of the earth ellipsoid's equator predicted in reference 1 and confirmed in the computer studies in reference 2, both studies made with a triaxial gravitational field. All the results of this study showing the dominance of the J_{22} (triaxial) term in the longitudinal perturbations, point strongly to the conclusion that such a circulation about the "stable" regions of zero longitude perturbation, exists for the full earth potential as well.

The latitude drift following perfect injection, predicted in (71–C) and, in fact for all injections where the initial latitude rate is zero, is initially a daily harmonic oscillation of small amplitude, but entirely on one side of the equator or the other depending on the sign of i which is proportional to the latitude perturbing force at the equator. Since J_{30} (the so-called "pear-shaped" harmonic of the earth's field) dominates i and it is negative, the perturbing force is directed south and the initial daily oscillation is southerly with an amplitude of $2i$. The orbital mechanics of this interesting effect is illustrated in figure 4. $\mathbf{F}_p$ is the southerly directed perturbing force causing a change in angular momentum of the satellite of $\delta\mathbf{P}_a$ from $\mathbf{P}_a\cdot\mathbf{r}$ is the radius vector to the satellite. It is seen from fig. 4.A, that if $\mathbf{E}_p$ acts south after an injection along the equator, the initial effect

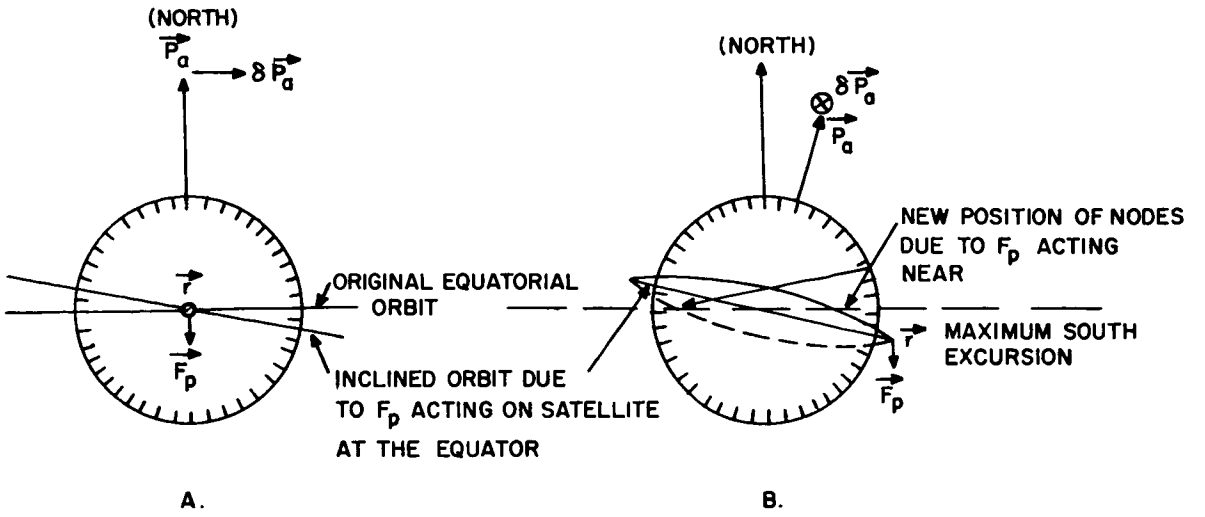


FIG. 4. Effect of latitude perturbation on a near equatorial orbit.

of $\mathbf{E}_p$ is to swing the satellite into an inclined orbit with the satellite moving south in this orbit. The injection longitude is at the descending node. Figure 4.B illustrates that when the satellite is 90° from its descending node, the effect of $\mathbf{F}_p$ is to *advance* the nodes and leave the inclination unchanged. At angles from the descending node between 0° and 90°, the effect is to advance the nodes to a lesser extent while increasing the inclination. From 90° to 180°, $\mathbf{F}_p$ advances the nodes to the same effect as in 0° to 90°, but the inclination change is reversed. Of course, during the southerly excursion, the earth's oblateness acts to regress the nodes. But the oblateness component normal to the orbit plane is extremely small and the linearized solution implies that for perfect injection into a very near equatorial orbit at near synchronous altitudes, the nodes actually advance about 180° per day; the latitude perturbation effect far overbalancing the oblateness. Furthermore, since the J_{30} perturbation acts in this manner for *all* near equatorial orbits, these orbits should show the same kind of effect; rapid nodal advance during the southerly swing, and regression during the northerly swing, both tending to keep the satellite longer to the south of the equator during every orbit regardless of altitude.

In summary; the approximate drift of a near equatorial, near synchronous satellite of the earth, injected at a radius where the radial perturbation force is near zero with near zero perturbation rates; is given by the following expressions;

$$r_1 \doteq \lambda_1 \doteq \phi_1 \doteq 0, \tag{72}$$

for injection at longitudes close to the four of (53) where the longitude perturbation force is zero;

$$r_1 \doteq +(2n)\,T + (-2n)\,\sin T,$$

at other longitudes; (73-A)

$$\lambda_1 \doteq 4n + (-4n)\,\cos T - \tfrac{3}{2}nT^2,$$

at other longitudes; (73-B)

$$\phi_1 \doteq i - (2ng)\,T - (i)\,\cos T + \tfrac{3}{2}(hn - ij)\,T^2,$$

at other longitudes. (73-C)

In (73-A) and (73-B), the products of perturbation constants (which would be of second order smallness) have been ignored (see app. A).

The drift of a near synchronous earth satellite injected with zero perturbation rates at −71.4° longitude from Greenwich (45° from the minor axis of the earth ellipsoid's equator, and thus having close to maximum longitude drift rates), as predicted by (73), is:

$$\Delta r = 259t - 41.25\,\sin(2\pi t) \text{ Feet.}$$

(t, in sidereal days). (74-A)

$$\Delta\lambda = 34.2 \times 10^{-6}(1 - \cos 2\pi t) - 505 \times 10^{-6} t^2 \text{ Degrees.}$$

(t, in sidereal days). (74-B)

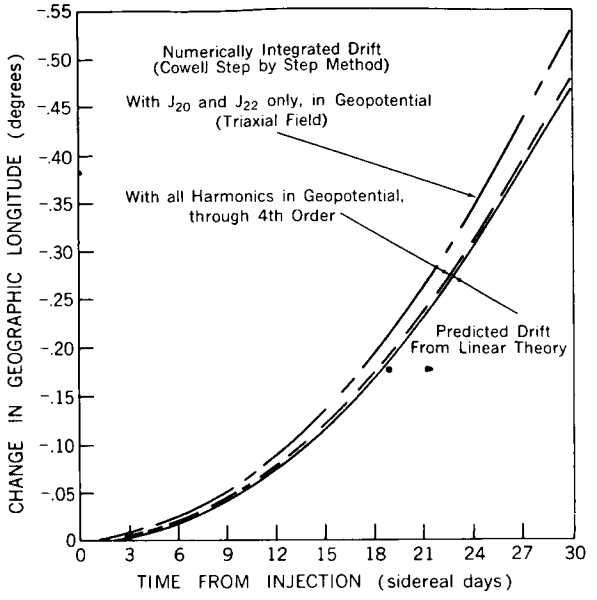


FIG. 5. Comparison of numerically integrated and predicted longitude drift of a near 24-hr satellite injected 45° east of the earth ellipsoid's minor equatorial axis with zero initial perturbation rates.

$$\Delta\phi = -10.7\times10^{-7}(1-\cos 2\pi t)+12.3\times10^{-12}t+5.64\times10^{-12}t^2 \text{ Degrees.}$$

(t, in sidereal days). (74–C)

Numerical integration (cowell step by step method) of the equations of motion have been carried out on an IBM 7090 computer and the comparison with the predictions of (74) is illustrated in figures 5 and 6 for the longitudinal and radial drifts to 30 days following injection. The full results indicate that the linear theory of (57) will predict the perturbations of a near synchronous, near equatorial satellite due to a "full" earth potential through 4th order, to within 2% in the mean daily longitude drift, and 1% in the mean daily radial drift for up to 180 days following a near perfect injection. The numerically computed latitude drift of the near synchronous satellite is of the order of $\pm10^{-6}$ degrees, maximum, over 180 sidereal days, which agrees with (74–C) in order of magnitude. Numerically integrated drifts for the above example in a "triaxial" (J_{20} and J_{22} harmonics only) earth field show errors of about 10% from the drift in a full potential field in both radius and longitude to 180 days following injection. This "error" was reduced to about 5% with the inclusion of the J_{31} harmonic in the programmed potential function. Inclusion of the J_{33} harmonic reduces the "error" to about 1% over 180 days.

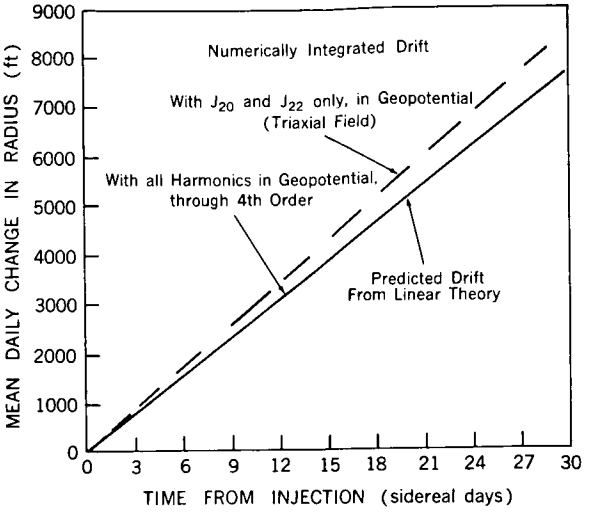


FIG. 6. Comparison of numerically integrated and predicted mean daily radial drift of a near 24-hr satellite injected 45° east of the earth ellipsoid's minor equatorial axis with zero initial perturbation rates.

It should be noted that the linear theory presented in this report does *not* predict the build up in eccentricity of the spiralling orbit of the resonant near synchronous satellite. However, the results of the numerical integration on the above example show that the *initial* eccentricity of the spiral orbit is well predicted. The radius in the numerically integrated orbit has a daily oscillation of about $\pm(34+5t)$ feet (t, in days) for a period of 5 days following injection; which is in excellent agreement with (74–A) considering the inherent machine error in the numerical integration.

CONCLUSIONS

The major conclusions of this paper may be summarized as follows:

1. There are four longitudes, located within $2\frac{1}{2}°$ of the axes fixed by the earth's elliptical equator, into which a 24-hour satellite may be placed and maintained with negligible drift for extended periods of time.

2. At intermediate longitudes (about 45° from

these "stable points") such a satellite will, even if injected "perfectly," experience a minimum of about 4° of drift in 3 months following injection.

3. Maximum drift of a perfectly injected 24-hour satellite in an earth gravity field to 4th order differs by about 10% from the drift experienced in a triaxial field.

4. Except in the immediate vicinity of the "stable points," the initial mean daily drift in radius following a perfect injection into a 24-hour orbit is proportional to time, and the initial mean daily drift in geographic longitude is essentially proportional to the square of time.

5. The maximum longitude perturbational force on a 24-hour satellite in an earth potential field to 4th order is approximately 12% greater than the maximum perturbation experienced in a triaxial field.

Appendix A

THE EVALUATION OF THE COEFFICIENTS C_{iq} IN THE APPROXIMATE INTEGRALS (57), FOR A COMPLETE SET OF INITIAL CONDITIONS s_0^0 AND s_0^1 ON THE DIMENSIONLESS PERTURBATIONS IN RADIUS, GEOGRAPHIC LONGITUDE AND LATITUDE

The zero'th through fifth derivatives of (57) at $T=0$ are given by:

$$\begin{aligned}
s_0^0\Delta q &= C_{1q} && +C_{5q} \\
s_0^1\Delta q &= && C_{2q}+C_{3q} && +C_{6q} \\
s_0^2\Delta q &= && +2C_{4q}-C_{5q} && -A_q. \\
s_0^3\Delta q &= && -C_{3q} && -3C_{6q} \\
s_0^4\Delta q &= && -4C_{4q}+C_{5q} \\
s_0^5\Delta q &= && +C_{3q} && +5C_{6q}
\end{aligned} \tag{A-1}$$

The solution of (A–1) for the coefficients C_{iq} in terms of the initial conditions s_0^0 through s_0^5 is:

$$\begin{aligned}
C_{1q} &= s_0^0\Delta_q+(s_0^4\Delta_q+2s_0^2\Delta_q+2A_q). \\
C_{2q} &= (s_0^1+s_0^5+2s_0^3)\Delta_q. \\
C_{3q} &= -\tfrac{1}{2}(3s_0^5+5s_0^3)\Delta_q. \\
C_{4q} &= -\tfrac{1}{2}(s_0^4\Delta_q+s_0^2\Delta_q+A_q). \\
C_{5q} &= -(s_0^4\Delta_q+2s_0^2\Delta_q+2A_q). \\
C_{6q} &= \tfrac{1}{2}(s_0^5+s_0^3)\Delta_q.
\end{aligned} \tag{A-2}$$

Differentiating (31)–(33) successively, the higher derivatives s_0^2 through s_0^5, in terms of s_0^0, s_0^1 and the perturbation constants (a, b, c, etc.) are:

$$s_0^2r_1=-as_0^0r_1-bs_0^1\lambda_1-cs_0^0\lambda_1-ds_0^0\phi_1+e.$$

$$s_0^2\lambda_1=-ks_0^1r_1-ls_0^0r_1-js_0^0\lambda_1-ms_0^0\phi_1+n.$$

$$s_0^2\phi_1=-gs_0^0r_1-hs_0^0\lambda_1-fs_0^0\phi_1+i.$$

$$s_0^3r_1=\{(-a+bk)s_0^1+lbs_0^0\}r_1+\{(-c)s_0^1+(bj)s_0^0\}\lambda_1+\{(-d)s_0^1+(bm)s_0^0\}\phi_1\cdots-bn.$$

$$s_0^3\lambda_1=\{(-l)s_0^1+(-ka)s_0^0\}r_1+\{(-j+kb)s_0^1+(kc)s_0^0\}\lambda_1+\{(-m)s_0^1+\cdots+(kd)s_0^0\}\phi_1+ke.$$

$$s_0^3\phi_1=(-g)s_0^1r_1-hs_0^1\lambda_1-fs_0^1\phi_1.$$

$$\begin{aligned}
s_0^4r_1=\{(bl+ck)s_0^1+(a^2-bka+cl+dg)s_0^0\}r_1+\{(ab+bj-b^2k)s_0^1+(ac-bkc\cdots+dh+cj)s_0^0\}\lambda_1 \\
+\{(bm)s_0^1+(ad-bdk+cm+df)s_0^0\}\phi_1-ae+bke-cn-di.
\end{aligned}$$

$$\begin{aligned}
s_0^4\lambda_1=\{(jk+ak-k^2b)s_0^1+(kl-klb+al+mg)s_0^0\}r_1+\{(j^2+ck+bl)s_0^1+(-kbj+\cdots lc+mh)s_0^0\}\lambda_1 \\
+\{(kd)s_0^1+(jm-kbm+mf+ld)s_0^0\}\phi_1+kbn-jn-el-mi.
\end{aligned}$$

$$s_0^4\phi_1=\{(kh)s_0^1+(fg+ag+lh)s_0^0\}r_1+\{(bg)s_0^1+(fh+cg+jh)s_0^0\}\lambda_1+\cdots$$
$$+\{(f^2+gd+mh)s_0^0\}\phi_1-fi-ge-nh.$$

$$s_0^5r_1=\{(a^2-bjk-bak+cl+dg+abk+b^2k^2)s_0^1+(-abl-bkl+b^2k-bal-\cdots bmg-cak)s_0^0\}r_1$$
$$+(ac-bj^2-bck-b^2l+cj-ckb+dh)s_0^1+\{(-abj+\cdots kb^2l-blc-bmh-c^2k)s_0^0\}\lambda_1$$
$$+\{(ad-kbd+cm+fd)s_0^1+(-abm-jbm-mbf+\cdots kb^2m-lbd-ckd)s_0^0\}\phi_1$$
$$+abn+jbn-kb^2n+ebl+mbi+cke.$$

$$s_0^5\lambda_1\{(jl-kbl-ck^2+al+mg-lbk)s_0^1+(-jka-a^2k+bk^2a-ckl-kdg-\cdots bl^2)s_0^0\}r_1$$
$$+\{(j^2-kbj-kab-kbj+b^2k^2+cl+mh)s^1+(-jkc-kac+\cdots bk^2c-kdh-ckj-blj)s_0^0\}\lambda_1$$
$$+\{(mj-kbm+ld+mf)s_0^1+(-kjd-kad+\cdots bdk^2-ckm-kdf-bml)s_0^0\}\phi_1$$
$$+jke+ake-bk^2e+ckn+dki+bnl.$$

$$s_0^5\phi_1=\{(gf+ag+hl-gbk)s_0^1+(-gal-kha)s_0^0\}r_1+\{(hf+cg+jh-kbh)s_0^1+\cdots(-bjg-kch)s_0^0\}\lambda_1$$
$$+\{(f^2+dg+mh)s_0^1+(-bmg-khd)s_0^0\}\phi_1+bng+keh. \quad \text{(A–3)}$$

(A–3) in (A–2) determines the C_{iq} of (57) for the "natural" initial value problem as;

$$C_{11}=\{[(bl+ck)]s_0^1+[1+(a^2-bka+cl+dg-2a)]s_0^0\}r_1$$
$$+\{[(ab+bj-b^2k-\cdots 2b)]s_0^1+[(ac-bkc+dh+cj-2c)]s_0^0\}\lambda_1$$
$$+\{(bm)s_0^1+[(ad-bdk+cm+\cdots df-2d)]s_0^0\}\phi_1+(-ae+bke-cn-di+2e+2A_1).$$

$$C_{12}=\{[(jk+ak-2k-k^2b)]s_0^1+[(kl-klb+al+mg-2l)]s_0^0\}r_1$$
$$+\{[(j^2+\cdots ck+bl)]s_0^1+[(1-kbj+lc+mh-2j)]s_0^0\}\lambda_1$$
$$+\{[(kd)]s_0^1+\cdots[(jm-kbm+ld+mf-2m)]s_0^0\}\phi_1+(-jn+kbn-el-mi+2n+2A_2). \quad \text{(A–4)}$$

$$C_{13}=\{[(kh)]s_0^1+[(fg+ag+lh-2g)]s_0^0\}r_1+\{(bg)s_0^1+[(fh+cg+\cdots jh-2h)]s_0^0\}\lambda_1$$
$$+\{[(1+f^2+gd+mh-2f)]s_0^0\}\phi_1+(-fi-ge-nh+2i+2A_3).$$

$$C_{21}=[(1+a^2-bjk-2bak+cl+dg+b^2k^2-2a+2bk)s_0^1+(-abl-bkl+\cdots b^2kl-cak-bal-bmg+2lb)s_0^0]r_1$$
$$+[(ac-bj^2+cj-bck-ckb-b^2l+\cdots dh-2c)s_0^1+(-abj+kb^2l-blc-bmh-c^2k+2bj)s_0^0]\lambda_1$$
$$+[(ad-kbd+cm+\cdots fd-2d)s_0^1+(-abm-jmb-mbf+kb^2m-lbd-ckd+2bm)s_0^0]\phi_1$$
$$+abn+jbn-\cdots kb^2n+ebl+mbi+cke-2bn.$$

$$C_{22}=[(jl-kbl-ck^2+al+mg-lbk-2l)s_0^1+(-jka-a^2k+bk^2a-ckl-kdg-\cdots bl^2+2ka)s_0^0]r_1$$
$$+[(1+j^2-kbj-kab-kbj+b^2k^2+cl+mh-\cdots 2j+2kb)s_0^1$$
$$+(-jkc-kac+bk^2c-kdh-ckj-bjl+2kc)s_0^0]\lambda_1$$
$$+\cdots[(mj-kbm+ld+mf-2m)s_0^0+(-kjd-kad+bdk^2-ckm-kdf-\cdots bml+2kd)s_0^0]\phi_1$$
$$+jkc+ake-bk^2e+ckn+dki+bnl-2ke.$$

$$C_{23}=[(gf+ag+hl-2g)s_0^1+(-gal-kha)s_0^0]r_1+[(hf+cg+jh-kbh-2h)s_0^1+\cdots(-bjg-kch)s_0^0]\lambda_1$$
$$+[(1+f^2+dg+mh-2f)s_0^1+(-bdg-khd)s_0^0]\phi_1+bng+keh.$$

$$C_{31}=\{[(-3/2)(a^2-bjk-2bak+cl+dg+b^2k^2)+(5a/2)-(5bk/2)]s_0^1$$
$$+\cdots[(-3/2)(-abl-bkl+b^2kl+cak-bal-bmg)-(5lb/2)]s_0^0\}r_1$$
$$+\cdots\{[(-3/2)(ac-bj^2+cj-ckb+dh-b^2l-bck)-(5c/2)]s_0^1$$
$$+\cdots[(-3/2)(-abj+kb^2l-blc-bmh-c^2k)-(5bj/2)]s_0^0\}\lambda_1$$
$$+\cdots\{[(-3/2)(ad-kbd+cm+fd)+(5d/2)]s_0^1$$
$$+[(-3/2)(-abm-ckd-jbm-\cdots mbf+kb^2m-lbd)-(5bm/2)]s_0^0\}\phi_1$$
$$+(-3/2)(abn+jbn-kb^2n+ebl+\cdots mbi+cke)+(5bn/2).$$

$$C_{32}=[(-3/2)(jl-kbl-ck^2-al+mg-lbk)+(5l/2)]s_0^1+$$
$$+\{[(-3/2)(-jka-a^2k+\cdots bk^2a-ckl-kdg-bl^2)+(5ka/2)]s_0^0\}r_1$$
$$+\{[(-3/2)(j^2-kbj-kab-kbj+\cdots b^2k^2+cl+mh)-(5/2)(-j+kb)]s_0^1$$
$$+(-3/2)(-jck-kac+bk^2c-kdh-\cdots ckj-bjl)-(5kc/2)]s_0^0\}\lambda_1$$
$$+\{[(-3/2)(mj-kbm+ld+mf)+(5m/2)]s_0^1$$
$$+\cdots[(-3/2)(-kjd-kad+bdk^2-ckm-bml-kdf)-(5kd/2)]s_0^0\}\phi_1$$
$$-\cdots(3/2)(jke+ake-bk^2e+ckn+dki+bnl)+5ke/2.$$

$$C_{33}=\{[(-3/2)(gf+ag+hl-gbk)+5g/2]s_0^1+[(-3/2)(-gal-kha)]s_0^0\}r_1$$
$$+\cdots\{[(-3/2)(hf+cg+jh-kbh)+5h/2]s_0^1+[(-3/2)(-bjg-kch)]s_0^0\}\lambda_1$$
$$+\{[(-\cdots 3/2)(f^2+dg+mh)+5f/2]s_0^1+[(-3/2)(-bmg-khd)]s_0^0\}\phi_1-(3/2)(bng+keh).$$

$$C_{41}=\{-(1/2)(bl+ck)s_0^1+(-1/2)(a^2+bka+cl+dg-a)s_0^0\}r_1$$
$$+\{(-1/2)(ab+bj-b^2k-\cdots b)s_0^1+(-1/2)(ac-bkc+dh+cj-c)s_0^0\}\lambda_1$$
$$+\{(-1/2)(bm)s_0^1+(-1/2)(ad-bdk+\cdots cm+df-d)s_0^0\}\phi_1+(1/2)(ae+bke-cn+di+e-A_1). \quad \text{(A–4)}$$

$$C_{42}=\{(-1/2)(jk+ak-k^2b-k)s_0^1+(-1/2)(kl-klb+al+mg-l)s_0^0\}r_1$$
$$+\{(-1/2)(j^2+\cdots ck+bl)s_0^1+(-1/2)(-kbj+lc+mh-j)s_0^0\}\lambda_1$$
$$+\{(-1/2)(kd)s_0^1+(-1/2)\cdots(jm-kbm+ld+mf-m)s_0^0\}\phi_1$$
$$+(-1/2)(-jn+kbn-el-mi+n-A_2).$$

$$C_{43}=\{(-1/2)(kh)s_0^1+(-1/2)(fg+ag+lh-g)s_0^0\}r_1$$
$$+[(-1/2)(bg)s_0^1+(-1/2)(fh+cg+\cdots jh-h)s_0^0]\lambda_1+[(-1/2)(f^2+gd+mh-f)s_0^0]\phi_1$$
$$+(-1/2)(-fi-ge-nh+i-A_3).$$

$$C_{51}=[(-1)(bl+ck)s_0^1+(-1)(a^2-bka+cl+dg-2a)s_0^0]r_1$$
$$+[(-1)(ab+bj-b^2k-\cdots 2b)s_0^1+(-1)(ac-bkc+dh+cj-2c)s_0^0]\lambda_1$$
$$+[(-1)(bm)s_0^1+(-1)\cdots(ad-bdk+cm+df)s_0^0]\phi_1+(-1)(ae+bke-cn-di+2e+2A_1).$$

Appendix B

THE EARTH GRAVITY POTENTIAL AND GRAVITATIONAL FORCE FIELD USED IN THIS REPORT

The gravity potential used is the exterior potential derived in reference 6 for geocentric spherical coordinates referenced to the earth's spin axis and its center of mass (see fig. 1). The harmonic series is truncated after J_{44}.

$$V_E=(\mu_E/r)\{1-(R_o/r)^2(J_{20}/2)(3\sin^2\phi-1)-3(R_o/r)^2J_{22}\cos^2\phi\cos2(\theta-\theta_{22})-\cdots$$
$$-(R_o/r)^3(J_{30}/2)(5\sin^3\phi-3\sin\phi)-(R_o/r)^3(J_{31}/2)\cos\phi[15\sin^2\phi-3]\cos(\theta-\cdots\theta_{31})$$
$$-(R_o/r)^315J_{32}\cos^2\phi\sin\phi\cos2(\theta-\theta_{32})-(R_o/r)^315J_{33}\cos^3\phi\cos3(\theta-\cdots\theta_{33})$$
$$+R_o/r)^4(J_{40}/8)(35\sin^4\phi-30\sin^2\phi+3)-(R_o/r)^4(J_{41}/8)\cos\phi[140\sin^3\phi-\cdots60\sin\phi]$$
$$\times\cos(\theta-\theta_{41})-(R_o/r)^4(J_{42}/8)\cos^2\phi[420\sin^2\phi-60]\cos2(\theta-\cdots\theta_{42})-(R_o/r)^4[J_{43}/8]$$
$$\times\cos^3\phi[840\sin\phi]\cos3(\theta-\theta_{43})-\cdots(R_o/r)^4(J_{44}/8)840\cos^4\phi\cos4(\theta-\theta_{44})\} \quad \text{(B-1)}$$

$$\mathbf{F}=\hat{r}F_r+\hat{\theta}F_\theta+\hat{\phi}F_\phi. \quad \text{(B-2)}$$

$$F_r=\dot{m}G_r=\dot{m}(\partial V_E/\partial r),$$

where;

$$G_r=(\mu_E/r^2)\{-1+(R_o/r)^2[1.5J_{20}(3\sin^2\phi-1)+9J_{22}\cos^2\phi\cos2(\theta-\theta_{22})+\cdots2(R_o/r)J_{30}(5\sin^2\phi-3)$$
$$\times(\sin\phi)+6(R_o/r)J_{31}(5\sin^2\phi-1)\cos\phi\cos(\theta-\cdots\theta_{31})+60(R_o/r)J_{32}\cos^2\phi\sin\phi\cos2(\theta-\theta_{32})$$
$$+60(R_o/r)J_{33}\cos^3\phi\cos3(\theta-\cdots\theta_{33})+0.625(R_o/r)^2J_{40}(35\sin^4\phi-30\sin^2\phi+3)-\cdots$$
$$+12.5(R_o/r)^2J_{41}(7\sin^2\phi-3)\cos\phi\sin\phi\cos(\theta-\theta_{41})+\cdots37.5(R_o/r)^2J_{42}(7\sin^2-1)$$
$$\times\cos^2\phi\cos2(\theta-\theta_{42})+\cdots525(R_o/r)^2J_{43}\cos^3\phi\sin\phi\cos3(\theta-\theta_{43})$$
$$+525(R_o/r)^2J_{44}\cos^4\phi\cos4(\theta-\theta_{44})]\}. \quad \text{(B-3)}$$

$$F_\theta=\dot{m}G_\theta=[\dot{m}/(r\cos\phi)](\partial V_E/\partial\theta),$$

where:

$$G_\theta=(\mu_E/r^2)(R_o/r)^2\{6J_{22}\cos\phi\sin2(\theta-\theta_{22})+1.5(R_o/r)J_{31}[5\sin^2\phi-1]\sin(\theta-\cdots\theta_{31})$$
$$+30(R_o/r)J_{32}\cos\phi\sin\phi\sin2(\theta-\theta_{32})+45(R_o/r)J_{33}\cos^2\phi\sin3(\theta-\cdots\theta_{33})+2.5(R_o/r)^2J_{41}$$
$$\times[7\sin^2\phi-3]\sin\phi\sin(\theta-\theta_{41})+15(R_o/r)^2(7\sin^2\phi-\cdots1)J_{42}\cos\phi\sin2(\theta-\theta_{42})$$
$$+315(R_o/r)^2J_{43}\cos^2\phi\sin\phi\sin3(\theta-\theta_{43})+\cdots420(R_o/r)^2J_{44}\cos^3\phi\sin4(\theta-\theta_{44})\}. \quad \text{(B-4)}$$

$$F_\phi=\dot{m}G_\phi=(\dot{m}/r)(\partial V_E/\partial\phi),$$

$$C_{52}=[(-1)(jk+ak-kb^2-2k)s_0^1+(-1)(kl-klb+al+mg-2l)s_0^0]r_1$$
$$+[(-1)(j^2+ck+\cdots bl)s_0^1+(-1)(-kbj+lc+mh-2j)s_0^0]\lambda_1$$
$$+[(-1=(kd)s_0^1+(-1)(jm+mf-kbm+\cdots ld-2m)s_0^0]\phi_1$$
$$+(-1)(-jn+kbn-el-mi+2n+2A_2).$$

$$C_{53}=[(-kh)s_0^1+(-1)(fg+ag+lh-2g)s_0^0]r_1+[(-bg)s_0^1+(-1)(fh+cg+jh-2h)s_0^0]\lambda_1$$
$$+\cdots[(-1)(f^2+gd+mh-2f)s_0^0]\phi_1+(-1)(-fi-ge-nh+2i+2A_3).$$

$$C_{61}=[1/2(a^2-bjk-2bak+cl+dg-a+bk+b^2k^2)s_0^1+1/2(-abl-bkl+b^2kl-bal-\cdots bmg-cak+lb)s_0^0]r_1$$
$$+[+1/2(ac-bj^2-bck-b^2l+cj-ckb+dh-c)s_0^1+\cdots 1/2(-abj-kb^2l-bcl-bmh-c^2k+bj)s_0^0]\lambda_1$$
$$+[+1/2(ad-kbd+cm+\cdots fd-d)s_0^1+1/2(-abm-jbm-mbf+kb^2m-lbd-ckd+bm)s_0^0]\phi_1$$
$$+1/2(abn+\cdots jbn-kb^2n+ebl+mbi+cke-bn).$$

$$C_{62}=[\tfrac{1}{2}(jl-kbl-ck^2+al+mg-l-lbk)s_0^1+\tfrac{1}{2}(-jka-a^2k+bk^2a-ckl-\cdots kdg-bl^2+ka)s_0^0]r_1$$
$$+[\tfrac{1}{2}(j^2-kbj-kab-kbj+b^2k^2+cl+mh-j+\cdots kb)s_0^1+\tfrac{1}{2}(-jkc-kac+bk^2c-kdh-ckj+kc-blj)s_0^0]\lambda_1$$
$$+[\tfrac{1}{2}(mj-kbm+\cdots ld+mf-m)s_0^1+\tfrac{1}{2}(-kjd-bml-kad+bdk^2-ckm+kd-bdl-\cdots kdf)s_0^0]\phi_1$$
$$+\tfrac{1}{2}(jke+ake-bk^2e+ckn+dki+bnl-ke).$$

$$C_{63}=[\tfrac{1}{2}(gf+ag+hl-g)s_0^1+\tfrac{1}{2}(-gal-kha)s_0^0]r_1+[\tfrac{1}{2}(hf+cg+jh-kbh-\cdots h)s_0^1+\tfrac{1}{2}(-bjg-kch)s_0^0]\lambda_1$$
$$+[\tfrac{1}{2}(f^2+dg+mh-f)s_0^1+\tfrac{1}{2}(-bmg-\cdots khd)s_0^0]\phi_1+\tfrac{1}{2}(bng+keh). \quad \text{(A-4)}$$

where,

$$G_\phi = (\mu_E/r^2)(R_\circ/r)^2\{-3J_{20}\sin\phi\cos\phi+6J_{22}\cos\phi\sin\phi\cos2(\theta-\theta_{22})-\cdots1.5(R_\circ/r)J_{30}(5\sin^2\phi-1)$$
$$\times\cos\phi+1.5(R_\circ/r)J_{31}(15\sin^2\phi-11)\sin\phi\cos(\theta-\cdots\theta_{31})+15(R_\circ/r)J_{32}(3\sin^2\phi-1)$$
$$\times\cos\phi\cos2(\theta-\theta_{32})+\cdots45(R_\circ/r)J_{33}\cos^2\phi\sin\phi\cos3(\theta-\theta_{33})+2.5(R_\circ/r)^2J_{40}(7\sin^2\phi-\cdots3)$$
$$\times\sin\phi\cos\phi+2.5(R_\circ/r)^2J_{41}(28\sin^4\phi-27\sin^2\phi+3)\cos(\theta-\cdots\theta_{41})+30(R_\circ/r)^2J_{42}$$
$$\times(7\sin^2\phi-4)\cos\phi\sin\phi\cos2(\theta-\theta_{42})+\cdots105(R_\circ/r)^2J_{43}(4\sin^2\phi-1)\cos^2\phi\cos3(\theta-\theta_{43})$$
$$+\cdots420(R_\circ/r)^2J_{44}\cos^3\phi\sin\phi\cos4(\theta-\theta_{44})\}. \quad \text{(B-5)}$$

The radial perturbation of the gravitational field referred to in this report is the residual of the sum of the gravitational and centrifugal forces on m at the moment of injection at the radius r_s. Since m is injected at an angular rate w for conditions where the initial perturbation rates with respect to the reference synchronous orbit are zero, the centrifugal force on m at the moment of this injection at r_s is; $\hat{r}(+w^2r_sm)$. Thus, the radial perturbation force, as defined above, at the moment of injection is: radial perturbation force $=mw^2r_s+mG_{r_s}=mw^2r_se$. Thus, the condition for the radial perturbation force (for a "perfect" injection, at injection) to be zero is: $e=0$.

Appendix C

THE MAGNITUDE OF THE LONGITUDE AND LATITUDE PERTURBATION FORCES PER UNIT MASS ON A NEAR SYNCHRONOUS SATELLITE, AROUND THE EQUATOR, DUE TO THE EARTH'S GRAVITY FIELD

It may be verified from appendix B, that:

1. The longitude perturbation force per unit mass on the near synchronous satellite is given by;

$$G_{\theta})_{r_s} \doteq (g_s)\,(a), \tag{C-1}$$

2. The latitude perturbation force per unit mass on the near synchronous satellite is given by;

$$(G_{\phi})_{r_s} \doteq (g_s)\,(i). \tag{C-2}$$

g_s is the radial acceleration of gravity in a $1/r^2$ earth field at the synchronous radius.

$$g_s = \mu_E/r_s^2 = [\mu_E/(R_\circ)^2](R_\circ/r_s)^2 = 32.15 \times 0.02288 = 0.735 \text{ ft sec}^2 \tag{C-3}$$

It is noted that at the synchronous radius;

$$\mu_E/w^2 r_s^3 \doteq 1. \tag{C-4}$$

In figure 7, the perturbation forces per unit mass are plotted for two earth gravity fields. One is for a field through 4th order due to Kozai (1962), and the other is for a simpler triaxial field which includes only the second order harmonic potential constants from the same source. It is seen that, while the triaxial components clearly dominate the longitude perturbation force function, the maximum force in the "full" field is 12% greater than the maximum force in the triaxial (table 1). The latitude perturbation in the triaxial field is, of course, zero. It is noted that the extreme magnitudes of the latitude perturbation function around the equator are about one order of magnitude less than the extremes of the

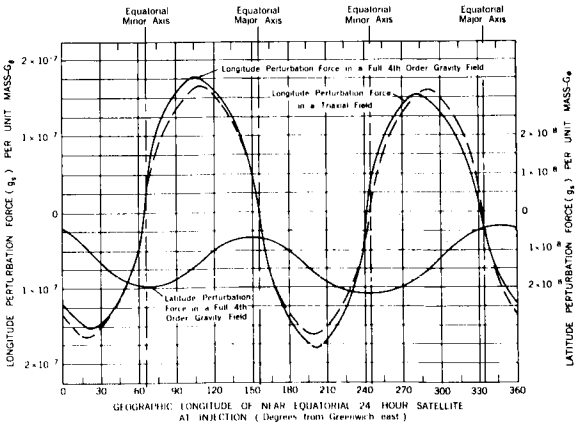


FIG. 7. The longitude and latitude perturbation forces on a near 24-hr satellite around the equator, due to the geopotential of Kozai (1962).

TABLE I

	Minimax $(G_\theta)_{r_s}$ (units of g_s)	λ_o (degrees)	% difference of minimax G_θ between full and triaxial field: (full-triax./traix.)×100
Triaxial field	-1.65×10^{-7}	18.6	+8.5
Full field	-1.51×10^{-7}	21.7	
Triaxial field	$+1.65\times10^{-7}$	108.6	+7.9
Full field	$+1.78\times10^{-7}$	109.7	
Triaxial field	-1.65×10^{-7}	198.9	−12.0
Full field	-1.85×10^{-7}	200.0	
Triaxial field	$+1.65\times10^{-7}$	288.9	−7.3
Full field	$+1.53\times10^{-7}$	282.3	

longitude perturbation function. This fact gives justification to the assumption that the regions of stability for the near synchronous satellite may be considered to be fixed by the longitude perturbation alone (i.e., by the zeros of n). It is also interesting that the minimum of the latitude perturbation in the full field at synchronous radius occurs close to the dynamically unstable region near the earth equator's major axis while the maximum latitude perturbation occurs near the dynamically stable region close to the minor axis. This coincidence tends to minimize the drift in the neighborhood of the major axis.

Appendix D

PROCEDURES FOR DETERMINING THE INJECTION RADIUS AND LONGITUDE INTO A NEAR SYNCHRONOUS ORBIT WITH MINIMAL INITIAL PERTURBATION ACCELERATIONS

The two conditions for a minimal drift, near synchronous earth orbit, are assumed to be:

$$e=0, \tag{D-1}$$

$$n=0. \tag{D-2}$$

These two conditions give a set of four injection radii and longitudes for minimum drift in a near synchronous equatorial earth orbit, with the potential of Kozai (1962). It can be assumed that injection at the earth rate into an inclined orbit whose nodes are near these geographic longitudes at near synchronous radii, will give orbits with similarly small nodal drift if the inclination is not excessive.

Let

$$r_s=r_o+\Delta r_s, \tag{D-3}$$

where $r_s \gg r_o$ is assumed and where r_o is the solution to:

$$w^2 r_0^3/\mu_E = 1+1.5J_{20}(R_o/r_o)^2-1.875J_{40}(R_o/r_o)^4 \tag{D-4}$$

r_o may be thought of as the synchronous radius for the "oblate" earth. With (D-3) and (D-4) in (D-1) and (D-2), ignoring all powers of $\Delta r_s/r_o$ greater than one and all terms in $J_{nm}\Delta r_s/r_o$; the synchronous conditions becomes:

$$\Delta r_s/r_o = \{(K_{31})_1 \cos\lambda_o+(K_{31})_2 \sin\lambda_o+[(K_{22})_1+(K_{42})_1]\cos 2\lambda_o+[(K_{22})_2+\cdots(K_{42})_2]\sin 2\lambda_o$$
$$+(K_{33})_1 \cos 3\lambda_o+(K_{33})_2 \sin 3\lambda_o+(K_{44})_1 \cos 4\lambda_0+\cdots(K_{44})_2 \sin 4\lambda_o\}/(3w^2 r_o^3/\mu_E), \tag{D-5}$$

and

$$0=(K_{31})_3 \sin\lambda_o+[(K_{22})_3+(K_{42})_3]\sin 2\lambda_o+(K_{33})_3 \sin 3\lambda_o+(K_{44})_3 \sin 4\lambda_o+\cdots(K_{31})_4 \cos\lambda_o$$
$$+[(K_{22})_4+(K_{42})_4]\cos 2\lambda_o+(K_{33})_4 \cos 3\lambda_o+\cdots(K_{44})_4 \sin 4\lambda_o. \tag{D6}$$

In (D-5) and (D-6), the gravitational constants K are;

$$(K_{31})_1=-6J_{31}(R_o/r_o)^3 \cos B_{31},$$
$$(K_{22})_1=-9J_{22}(R_o/r_o)^2 \cos 2B_{22},$$
$$(K_{42})_1=37.5J_{42}(R_o/r_o)^4 \cos 2B_{42},$$
$$(K_{33})_1=-60J_{33}(R_o/r_o)^3 \cos 3B_{33},$$
$$(K_{44})_1=-525J_{44}(R_o/r_o)^4 \cos 4B_{44},$$
$$(K_{31})_3=-1.5J_{31}(R^o/r_o)\cos B_{31},$$
$$(K_{22})_3=6J_{22}\cos 2B_{22},$$
$$(K_{42})_3=-15J_{42}(R_o/r_o)^2 \cos 2B_{42},$$
$$(K_{33})_3=45J_{33}(R_o/r_o)\cos 3B_{33},$$
$$(K_{44})_3=420J_{44}(R_o/r_o)^2 \cos 4B_{44}, \tag{D-7}$$

where

$$(K_{nm})_2 = (K_{nm})_1 \tan(nB_{nm}), \quad \text{and;} \quad (K_{nm})_4 = -(K_{nm})_3 \tan(nB_{nm}).$$

To solve (D–4) for the oblate earth synchronous radius r_o, let;

$$r_o = r_{01} + \Delta r_o, \tag{D–8}$$

where r_{01} is the "spherical earth" synchronous radius satisfying;

$$w^2 r_{01}^3/\mu_E = 1. \tag{D–9}$$

If $\Delta r_o \ll r_{01}$ is assumed, then with (D–9) and (D–8) in (D–4), ignoring the $J_{nm}\Delta r_o/r_{01}$ as of second order, the oblate synchronous radius is determined from;

$$\Delta r_o/r_{01} = (R_o/r_{01})^2[1.5J_{20} - 1.875J_{40}(R_o/r_{01})^2]3. \tag{D–10}$$

The procedure to determine the elements of the "stable" synchronous orbit as outlined on page 508 is as follows (the results for the potential due to Kozai (1962) appearing on pages 508 and 509)

1. Solve for the spherical synchronous radius r_{01} from (D–9).
2. Solve for the oblate synchronous radius r° from the results of the solution for Δr° from (D–10) in (D–9).
3. Solve for the constants K in (D–7).
4. Solve for the "stable" synchronous injection or nodal longitudes λ_o from (D–6).
5. Solve for the "stable" synchronous radii r_s corresponding to these longitudes from the results of the solution for Δr_s from (D–5) in (D–3).

For the earth potential used, the magnitude assumptions in D–1 and D–2 are valid and the "uncoupled" procedure 1 through 6 above establishes the "stable" synchronous elements to high accuracy.

REFERENCES

1. Frick and Garber, Perturbations of a Synchronous Satellite Due to the Triaxiality of the Earth, Rand Corp. Memo. RM–2996–NASA, Jan. 1962.
2. Blitzer, Boughton, Kang, and Page, Effect of Ellipticity on 24-Hour Nearly Circular Satellite Orbits, *J. Geophys. Research*, **67**, 329–335 (1962).
3. Barrett, The Perturbations of a Synchronous Satellite Resulting from the Gravitational Field of a Triaxial Earth. (Mission Analysis Report MAR 2 of the Advanced Missions and Research Section, Spacecraft Systems and Projects Division of GSFC NASA GSFC Gray Cover Document X–623–62–160 (1962).
4. Kozai, A Recent (1962) Determination of the Harmonic Coefficients in the Earth's Gravity Potential as Derived by Y. Kozai of the Smithsonian Astrophysical Observatory (in Memo. 623.12(62)206; to W. C. Isley, Spacecraft Systems and Projects Division of GSFC NASA, from C. A. Wagner, Nov. 9, 1962).
5. Isley, A Summary of Constants Associated with Orbital Analysis of Earth Satellites Including the Influence of Their Uncertainties Upon Gravitational Measurements for Synchronous Satellites (GSFC NASA Gray Cover Document X–623–62–169, 1962).
6. Wagner, The Gravitational Potential of a Triaxial Earth (GSFC NASA Gray Cover Document X–623–62–206, 1962).

III SOLAR PHYSICS

PRELIMINARY SOLAR FLARE OBSERVATIONS WITH THE SOFT X-RAY SPECTROMETER ON ORBITING SOLAR OBSERVATORY I

WILLIAM E. BEHRING, WERNER M. NEUPERT, AND JOHN C. LINDSAY

Goddard Space Flight Center

A spectrometer carried by Orbiting Solar Observatory I (1962 ζ 1) and used for observing the solar spectrum from 10 to 400A is described. The instrument utilizes a concave grating mounted in grazing incidence, and an open-window photomultiplier. Resolution of approximately 1A is obtained throughout the wavelength region covered. Solar spectra obtained with this instrument during a rocket flight are shown, and tentative identification of the origins of observed spectral lines is listed. Preliminary satellite results obtained during a solar flare are discussed.

INTRODUCTION

The primary experiment on Orbiting Solar Observatory I (1962 ζ1) was a soft X-ray spectrometer designed specifically to make satellite measurements of the solar spectrum in the wavelength range from 109 to 400 A. As a result of the successful launching of the satellite into a nearly circular earth orbit (550 km perigee and 600 km apogee) and the subsequent successful operation of the experiment, the first long-term measurements of the soft X-ray solar spectrum have been obtained. The purpose of this paper is to briefly describe the instrument, to illustrate its performance with data obtained from a rocket launch, and to present some preliminary satellite data obtained before and during the solar flare of March 13, 1962.

THE SPECTROMETER

The angular aperture of the spectrometer is approximately 1.2° by 2.2°. Hence, with moderately accurate pointing (within a few angular minutes of the center of the solar disk) the spectrometer responds to the total light intensity emitted by the sun. The instrument's orientation is such that sunlight is incident perpendicularly on the front face, passes through the entrance slit, and strikes a concave grating mounted in grazing incidence, the angle of incidence being 88° (fig. 1). The grating, an original which was lightly ruled in a special glass by the Nobel Institute in Stockholm, has 576 lines/mm on a blank with a radius of curvature of 1 m. No reflecting coatings

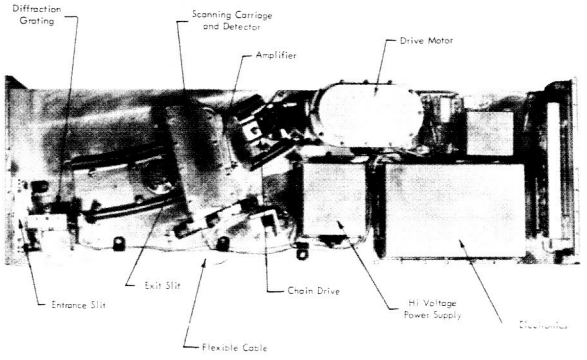


FIGURE 1.—OSO I Spectrometer.

were used on this grating. The diffracted rays continue on to the exit slit in front of the detector. The detector is mounted on a carriage which is driven on a circular rail so that the exit slit scans along the Rowland circle from 10 to 400 A. The plane of the exit slit stays approximately perpendicular to the diffracted ray at all positions along the rail, thereby keeping the spectral

passband nearly constant for all angles of diffraction. The 50-micron entrance and exit slits provide a spectral passband of 1.7A and permit resolution of lines 0.85A apart.

The detector used was an open-window multiplier phototube developed by the Bendix Corp. specifically for use in this spectrometer and now commercially designated M–306. Photoelectrons from the tungsten cathode move along cycloidal paths in crossed electric and magnetic fields between two glass strips, each coated with a semiconducting secondary-emitting oxide layer. One of the glass strips serves as a continuous dynode. Each photoelectron is multiplied into a pulse of approximately 10^6 electrons at the anode. These electrical pulses are amplified and, after coding to compress bandwidth, are recorded on a tape recorder for later transmission to a ground station.

The spectrometer uses about 1.3 w supplied by the satellite at 18 v dc. About 1 w of this goes to the oscillator powering the three-phase synchronous motor, which requires about 300 mw at 137 cps to yield 100 mw of mechanical output power. The remainder of the power is used in the multiplier and pulse handling circuitry.

All of the materials exposed inside the spectrometer were tested at a pressure of about 10^{-6} mm Hg in order to eliminate any material that had a high vapor pressure. Because the electronic circuits were potted in a compound having a high vapor pressure, they were enclosed in sealed boxes vented to the outside through holes in the spectrometer base plate. During assembly all the parts exposed to the interior were carefully cleaned so that they would be free of oil, grease, and other contaminants. The temperature of the instrument was controlled by the radiation balance. Most of the outer case was polished. The central section was painted black in order to keep the temperature close to that of the satellite.

Alinement of the spectrometer was accomplished with the use of only visible light by means of a fixture with a radius rod pivoted at the center of the Rowland circle; this procedure was developed on the basis of the method described by Rathenau and Peerlkamp.[1] Alinement tests were performed with a source of carbon K radiation. A separate photomultiplier was used to provide a monitor on the stability of this source. The response to scattered hydrogen Lyman-alpha radiation was checked with a closed hydrogen discharge lamp. The specific grating used was selected by means of comparative tests performed on several gratings.

THE ROCKET FLIGHT

Figure 2 shows the solar spectrum obtained with a similar instrument and a rocket pointing control during an Aerobee rocket flight. In this spectrum the wavelength regions of 120 to 170 A and 220 to 240A represent the average of data taken in three different scans in an attempt to improve the reliability of faint lines and provide continuity in the region originally containing wavelength marker pulses. In the region below 100A evidence of spectral lines is inconclusive. Comparison of the observed counting rates with laboratory scattered-light measurements indicates that for the rocket flight the signal attributable to the first order spectrum becomes lost in the scattered light below about 60A.

An attempt has been made to identify the resonance lines of highly ionized atoms of the heavier elements. Lines produced by several stages of ionization of C, N, and O, and also by Mg, Si, Ne, and Fe have already been found at wavelengths greater than 400A. The extension of isoelectronic sequences to heavier elements leads to resonance lines with wavelengths below 400A. The wavelengths of some of the strong lines observed are shown in table 1, along with tentative identification of their origin. The tabulation

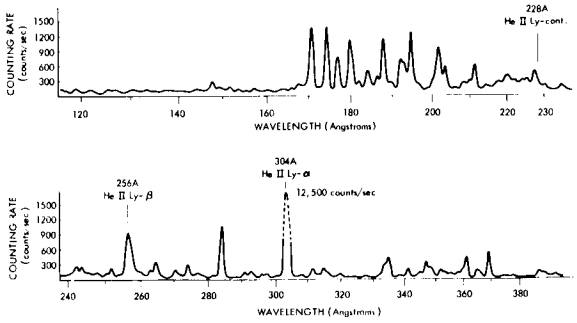


FIGURE 2.—The solar spectrum from 120 to 400 A as recorded near peak altitude (201–216 km) during an Aerobee rocket flight at 1433 GMT on September 30, 1961.

TABLE 1.—*Tentative Identification of Various Strong Emission Lines in the Solar Spectrum*

Element	Wavelength (angstroms)
Fe XIV, Fe XII	370
Fe X	366
	365
Fe XVI	361
Fe XII	360
Ni XV, Fe X	347
Fe XIV	345
	341
Fe XVI	335
Ni XV	333
	320
Ni XV	316
	315
	312
He II	304
	293
Cu XIX	274
	271
S X	264
A XIV	263
He II, Ni XVII	256
A XIV	250
He II	244
	243
He II	234
Ca XV	227
S IX	220
	211
	204
	202
	195
	193
	192
A XI	188
A XI	186
Cl IX	184
	182
Cl IX	180
	177
	174
	171

of emissions compiled by Varsavsky[2] was used in this work. The criteria for assigning values were: (1) Agreement with theoretically extrapolated values of the spacing and relative intensities of members of a multiplet, with the assumption, for the intensities, of an optically thin corona; (2) approximate agreement in wavelength with theory for lines not yet observed under laboratory conditions; (3) observation of more than one stage of ionization.

A preliminary analysis of the spectrum was made for ions known to exist in the solar atmosphere. These are ions of Fe, Ni, and, with lesser abundances, Ca and A. Identification of iron multiplets on the basis of one observation is difficult because of the presence of strong second-order lines as well as the superposition of the multiplets themselves. Only the Fe XV line has been calculated with accuracy (by Edlen—see reference 3), and it has been identified with a strong line at 284A. The resonance lines of Ca XII and Ca XIII (two ions observed in the visible coronal spectrum) cannot be associated with any of the emissions in the far ultraviolet spectrum. Nor is a correlation observed, although it is expected, for A X.

All of the foregoing assignments must be regarded as tentative and are presented as "work in progress." It is expected that it will be possible to combine the application of satellite observations with further theoretical analyses to achieve more reliable identifications.

Long term observations from OSO I will permit additional conditions to be applied: (1) Constancy in time of the ratio of intensities of the members of a multiplet, under the assumption that the opacity of the corona does not change with time; (2) regularity in the variations of intensity throughout the observed stages of ionizations; (3) for each stage of ionization, agreement of intensity variation with the corresponding variation in the visual coronal line(s) for that stage.

PRELIMINARY SATELLITE OBSERVATIONS

The Orbiting Solar Observatory I was launched at 1606 U.T. on March 7, 1962, during a period of low solar activity. The first flares observed after launch were in McMath plage 6366, with heliographic coordinates N10E66. The flares were of importance 1 and 2+ and occurred at 1407 and 1444 U.T., respectively. Figure 3 is an H-alpha photograph of the sun showing the 2+ flare as it appeared at 1543:22, considerably after maximum (which occurred at about 1450 U.T.). The total duration of this flare was about 2 hours. Ionospheric effects were recorded during the entire

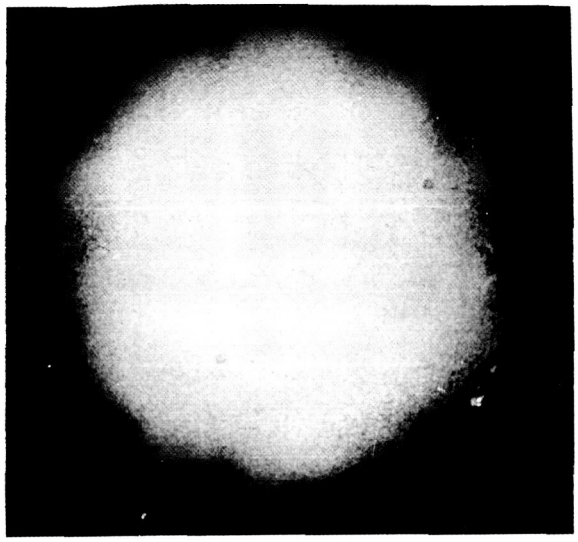


FIGURE 3.—Picture of H-alpha flare at 1453:22, March 13, 1962; picture presented through the courtesy of the High Altitude Observatory (HAO), Boulder, Colo.

life of the visible flare, and included an SWF (short-wave fadeout) of importance 3, a 99% SPA (sudden phase anomaly), SCNA (sudden cosmic noise absorption) of importance 1 (20%), and SEA (sudden enhancement of atmospherics) of importance 2. Type III spectral radio bursts were observed on 21–41 Mc from 1514 U.T. to 1632 U.T. and continuum bursts from 1519 to 1550 U.T. Major bursts were also observed on the single frequencies of 108 and 2800 Mc. There was no evidence of energetic particle fluxes.[4]

In the time available, it has been possible to reduce only partially the satellite data for the six orbits encompassing the visual flare. Two spectral lines were chosen for analysis: He II Lyman-alpha at 304A; and a line at 284A (tentatively identified as an Fe XV line). These data are shown in figure 4, from which we observe two points.

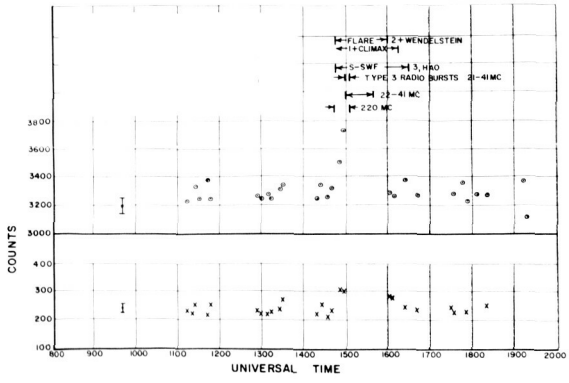


FIGURE 4.—Relative solar flux at 304 A (middle) and 284 A (bottom) during the March 13, 1962, flare (Wendelstein and Climax are observatories).

1. Increased emission coinciding with or preceding visual observations or other indicators.
2. Possible continued enhancement after the cessation of other indicators.

Using values for these two lines before the onset of the flare and during the flare, we find the maximum enhancements for the lines were 15% for 304A and 28% for 284A. Typical error flags are shown to the left on the figure. Practically all the error is statistical, and is due to the relatively small numbers of photons counted. The changes in the observed line intensities are larger than the expected errors and are believed to represent real changes associated with the flare.

CONCLUSIONS

As of April 13, 1962, OSO I had made approximately 550 orbits of the earth, and 3500 spectra of the sun in the 10–400A region had been collected. During this time 14 flares of varying importance occurred—the largest one, of importance 3, on March 22. It is believed that the OSO data will:

1. Aid in identifying the spectral lines;
2. Allow limits to be placed on the continua in this wavelength region, for a quiet sun as well as during solar activity;
3. Allow quiet-sun line intensity measurements to be made with some certainty;
4. Determine line intensity enhancement, if any, before, during, and after visual flare activity;
5. Determine enhancement, if any, associated with other solar activities such as plage areas and spot groups;
6. Determine short time fluctuations in line intensities not associated with other easily observable phenomena.

Analysis and interpretation of these results should throw light on the energy transport in the corona and the relaxation time of the corona, as

well as form the basis for a more complete model of the chromosphere.

ACKNOWLEDGMENTS

Many people played an important part in the development of this spectrometer. The authors would like to thank especially Prof. D. H. Tomboulian of Cornell University, who developed and carried out the tests on the basis of which the flight gratings were selected. He also gave the benefit of his counsel and his knowledge of instruments and phenomena in the soft X-ray and vacuum ultraviolet regions. Mr. Kennard Saffer and Mr. Paul Kraft of the U.S. Naval Weapons Plant in Washington, D. C., designed and supervised the construction of the spectrometer. Mr. W. A. Nichols of Goddard Space Flight Center carried out the modifications required for the rocket spectrometer and assisted in the development of the satellite spectrometer.

REFERENCES

1. RATHENAU, G., and PEERLKAMP, P. K., "Zur Justierung des Konkavgitters in streifender Inzidenz," *Physica*, vol. 2, pt. 1, pp. 125–143, 1935.
2. VARSAVSKY, C. M., "Some Atomic Parameters for Ultraviolet Lines," *Astrophys. J. Suppl. Ser.*, **6**(53): 75–108, March 1961.
3. FIROR, J., and ZIRIN, H., "Observations of Five Ionization Stages of Iron in the Solar Corona," *Astrophys. J.*, **135**(1): 122–137, January 1962.
4. "CRPL–F, Part B, Solar-Geophysical Data," issued monthly by the Nat. Bur. Standards, Central Radio Propagation Lab., Boulder, Colo.

THE EFFECT OF REFRACTION ON THE SETTING SUN AS SEEN FROM SPACE IN THEORY AND OBSERVATION

WINIFRED SAWTELL CAMERON

Goddard Space Flight Center

LT. COL. JOHN H. GLENN

U.S. Marine Corps

LT. CDR. M. SCOTT CARPENTER

U.S. Navy

JOHN A. O'KEEFE

Goddard Space Flight Center

The theory of refraction predicts that the setting sun or moon as seen from space should be highly flattened. The Mercury Project Manned Space Flights MA-6 and MA-7 have provided photographs of the phenomenon. To compare theory with observation, B. Garfinkel's Method I of computing refraction for large zenith distances was used to construct theoretical solar profiles for four true zenith distances of the center of the setting sun for comparison with the photographs taken by Glenn and Carpenter.

INTRODUCTION

The problem of the refraction of light by the earth's atmosphere as seen from a space capsule differs essentially from the problem as seen from the surface of the earth. At the earth's surface it is possible to calculate the astronomical refraction within 1 second of arc by Comstock's formula down to elevation angles of 15° above the horizon. Comstock's formula which depends only on the elevation angle and the index of refraction at the observer, would be the same if the earth were flat and surrounded by a 100-ft atmosphere. Near the horizon, it is true, terms involving the scale height and the curvature of the earth must be introduced. The problem of the lateral displacement of the ray by refraction is hardly considered, except in certain eclipse calculations.

In the case of the capsule, on the other hand, owing to the great distance (of the order of 1000 km) from the observer to the relevant region of the atmosphere, the variation of the refractive index with path is an essential part of the computation. We cannot approach the problem without a good knowledge of the scale height and of the curvature of the earth. The lateral displacement of the ray is relatively enormous.

The observation of the rising and setting of the sun in Mercury project manned orbital flights has emphasized the need for a more complete theory. The solar image should appear strongly flattened—almost sausage-shaped. Astronauts Glenn and Carpenter obtained photographs of the setting sun that illustrate this rather striking effect.

A THEORETICAL SOLAR PROFILE

The general procedure for computing refraction at extreme altitudes is presented in order to

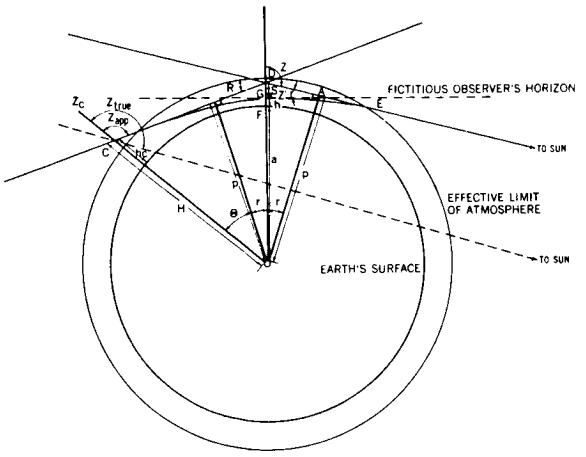


FIGURE 1.—Geometry of a ray from the setting sun as seen from the capsule.

construct a theoretical solar profile for comparison with photographic data. The procedure is applied to Carpenter's orbital conditions on May 24, 1962. The quantities to be determined are the apparent zenith distance Z_{app} and the true zenith distance Z_{true}, as seen from the capsule. To find these a ray through the atmosphere to the capsule is idealized. The phenomenon occurs effectively only for rays whose perigees are less than 20 km above the earth's surface. Thus we consider rays at 2 km intervals up to an altitude h of 20 km.

In figure 1, the ray from the sun is traced backward from the capsule C. In the first section, from the spacecraft to the atmosphere, X is straight. If the ray continued in this direction toward the sun, there would be a point B of nearest approach to the center of the earth O. This distance is denoted by p, and the angle at the center of the earth from the capsule to B by Θ. If B and p are known, the apparent height of any point on the sun (as seen from the spacecraft) could be calculated. To make the calculation, the curving optical ray is followed forward until it is refracted so as to be parallel to the surface of the earth. This point is called the perigee of the ray, and is denoted by G. The line OG makes an angle $\Theta+r$ with OC, where r is the refraction angle for the sun when it is seen 90° from the zenith by an imaginary observer at G.

If the straight portion of the ray is prolonged, it will intersect OG at some point D. Then the height of D above G is called the refractive height s. For any given height, say G, the refraction angle r at the horizon and the refractive height s which depends on the true height and r can be calculated. Then the right triangle OBD can be solved for the distance p from the center of the earth to the straight line prolongation of the space portion of the ray C. The length p is denoted, by analogy with the similar dynamical problem, as the impact parameter.

Thus given p and the capsule height, the apparent angles at the spacecraft can be calculated as a function of Θ. The refraction angle $R=2r$ is added to Θ to form the true zenith distance.

The computation of the refraction $r=z-z'$, where z is the true zenith distance and z' the apparent zenith distance, for the imaginary observer stationed at perigee, was based on the rather detailed theory of B. Garfinkel (method I).[1] The pertinent formulas are:

$$r = T^{1/2} \sum_{i=0}^{5} B_i W^{i+1} ,$$

$$\cot \theta = \gamma T^{-1/2} \cot z ,$$

$$W = PT^{-2} ,$$

where z is 90°, T the absolute temperature at h divided by 273°.0, P the pressure at h divided by the ground pressure of 1.013×10^6 dynes/cm^2, B a coefficient involving the index of refraction μ and the polytropic index n (for $z=90°$: $B_0=2012''.2$, $B_1=168''.2$, $B_2=21''.8$, $B'=3''.2$, $B'=0''.5$), and γ a constant dependent upon n.

The temperature, pressure and density (δ) of the atmosphere at an altitude h were taken from the Rocket Panel data.[2] More recent data are available from CIRA,[3] but the results are not significantly different for this computation.

For greater accuracy than required here, corrections to the approximations of Garfinkel's method may be made with the formula:

$$E = \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 = F_1 pT^{-1/2} + \frac{F_2 hR}{10^4} + \epsilon_3 + T^{1/2} \csc\theta \sum_{i=0}^{\infty} J_i \left(w \tan^2 \frac{\theta}{2}\right)^{i+1} \delta n ,$$

TABLE 1.—*Data Obtained in the Computation of the Apparent and True Zenith Distances Using the Final Definitive Elements of the MA–7 Orbit*

Altitude h (meters)	Temperature ÷ 273.0 °K	Pressure ÷ 1.013×10^6 dynes/cm²	Ratio of densities* δ	Refraction r (minutes of arc)	Index of refraction μ	Refractive height s (meters)	Impact parameter p (meters)	Θ (degrees)	Apparent zenith distance Z_{app} (degrees)	True zenith distance Z_{true} (degrees)
0000	1.0000	1.0000	1.0000	36.765	1.0002944	2240.1	6,372,896	15.949	105.949	107.175
2000	1.0330	0.7932	0.8532	27.081	1.0002512	1797.9	6,374,620	15.894	105.894	106.797
4000	0.9985	0.6214	0.6903	22.073	1.0002032	1425.8	6,376,314	15.841	105.841	106.577
6000	0.9524	0.4812	0.5611	18.188	1.0001652	1141.7	6,378,072	15.785	105.785	106.391
8000	0.8974	0.3676	0.4551	15.092	1.0001340	915.5	6,379,873	15.728	105.728	106.231
10,000	0.8454	0.2757	0.3623	12.299	1.0001067	720.6	6,381,699	15.669	105.669	106.079
12,000	0.8040	0.2038	0.2819	9.742	1.0000830	554.3	6,383,548	15.610	105.610	105.935
14,000	0.7751	0.1488	0.2128	7.468	1.0000626	413.5	6,385,419	15.550	105.550	105.799
16,000	0.7619	0.1075	0.1567	5.508	1.0000461	302.0	6,387,313	15.489	105.489	105.673
18,000	0.7656	0.0775	0.1124	3.922	1.0000331	214.7	6,389,231	15.426	105.426	105.557
20,000	0.7795	0.0562	0.0802	2.757	1.0000236	152.3	6,391,170	15.363	105.363	105.455

*δ is the density at h divided by the density at the earth's surface (1.72×10^{-3} gm/cm³) and is tabulated for the computation of μ.

where F_1, F_2, J_0, J_1, J_2 are tabulated as functions of θ, and ϵ' is tabulated as a function of $|90°-z|$ and h. Note that a factor of 10^4 was omitted in Garfinkel's paper and ϵ_2 should be computed according to the above equation.

The parameter s, which we called the refractive height, is a refraction correction commonly applied in calculations of times of contact in eclipses. The derivation of s is to be found in Chauvenet.[4] His equation 564 gives its relation to the index of refraction as $1+s/a=\mu(\sin z')/\sin z$, where a is the mean radius of the earth (6,371,020m), μ the index of refraction at h, z' the apparent zenith distance (90°) for a fictitious observer at G, and z the true zenith distance $(z'+r)$ at the same point.

When μ, r, and s have been obtained, p is obtained from the equation $p=(a+h+s)\cos r$. Θ is then determined from the relation $\cos\Theta=p/H$, where $H=a+h_c'$ $h_c=257{,}000m$ as determined by the orbit computed from the final definitive elements of Carpenter's orbit. Finally, Z_{app} and Z_{true} are related to Θ and R such that $Z_{app}=90°+\Theta$ and $Z_{true}=90°+(\Theta+R)$. The computed results are summarized in table 1.

The flattening of the image of the setting sun may be illustrated by plotting Z_{app} against Z_{true}. An image representing the sun (to scale) may be placed at any Z_{true}, and points around the limb, extended to the curve, may be located on the Z_{app} axis, thereby giving the apparent zenith distance of each point. Since the horizontal axis is not affected by refraction, parallels of altitude (almucantars) may be laid off on the unrefracted image of the sun, and similarly laid off on the apparent image of the sun. The latter may be rectified for easy comparison. The theoretical profiles of four phases of a setting sun are illustrated in figure 2. Here Z_{true} versus Z_{app} is plotted with images of the sun given for four true zenith distances of the sun's center: (fig. 2a) $Z_{true}=105°.455$; (fig. 2b) $Z_{true}=106°.231$; (fig. 2c) $Z_{true}=106°.915$ (sun's lower limb on the horizon) and (fig. 2d) $Z_{true}=107°.175$ (sun's center on horizon). The ratio of the vertical to horizontal diameters are approximately 0.63, 0.36, 0.17, and 0.11 respectively. Considering the capsule angular velocity (4°/min), we see that the entire effect treated herein would occur for the astronaut during a relatively short time interval (about 20 sec).

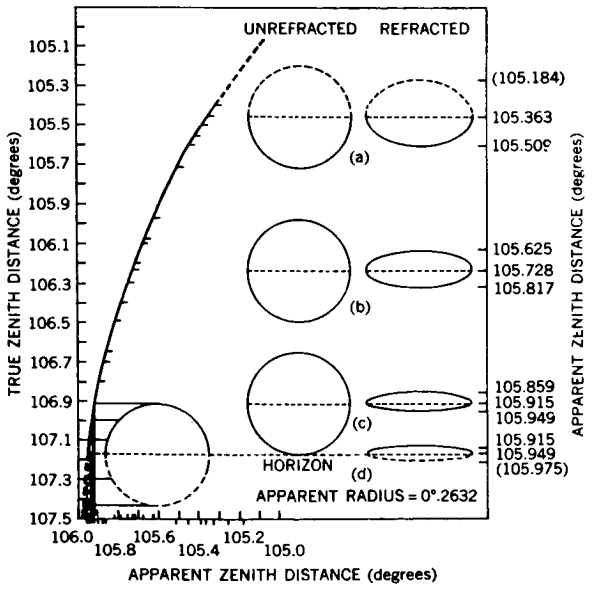


FIGURE 2.—Stages of the setting sun for four zenith distances.

COMPARISON OF PHOTOGRAPHIC AND COMPUTED DATA

Uncertainty of time in the photographic record precludes an exact comparison of theory and observations. However, figure 2c perhaps most nearly simulates the photographs (fig. 3 and 4), which show the effects of the capsule's motion—somewhat increased vertical diameter and some lateral motion—of the image but still demonstrate the effect. Figure 3 was photographed on the MA–6 orbital flight of February 20, 1962. The sun was not seen then as a narrow, flat object, but instead was seen to spread out about 10° on either side and to merge with the twilight band. In the original photograph the true setting sun and horizon appear with reflections both above and below. The phenomenon is perhaps more clearly seen in the lower reflection, as shown in figure 3.

Figure 4 was photographed on the MA–7 orbital flight on May 24, 1962. At that time the sun was observed to be definitely flattened during sunrise and sunset and very similar to its appearance in the photographs. The flattening effect of refraction on a setting celestial object as seen above the atmosphere—a condition simulated by

FIGURE 3.—Photograph of setting sun taken by Glenn on the MA–6 orbital flight.

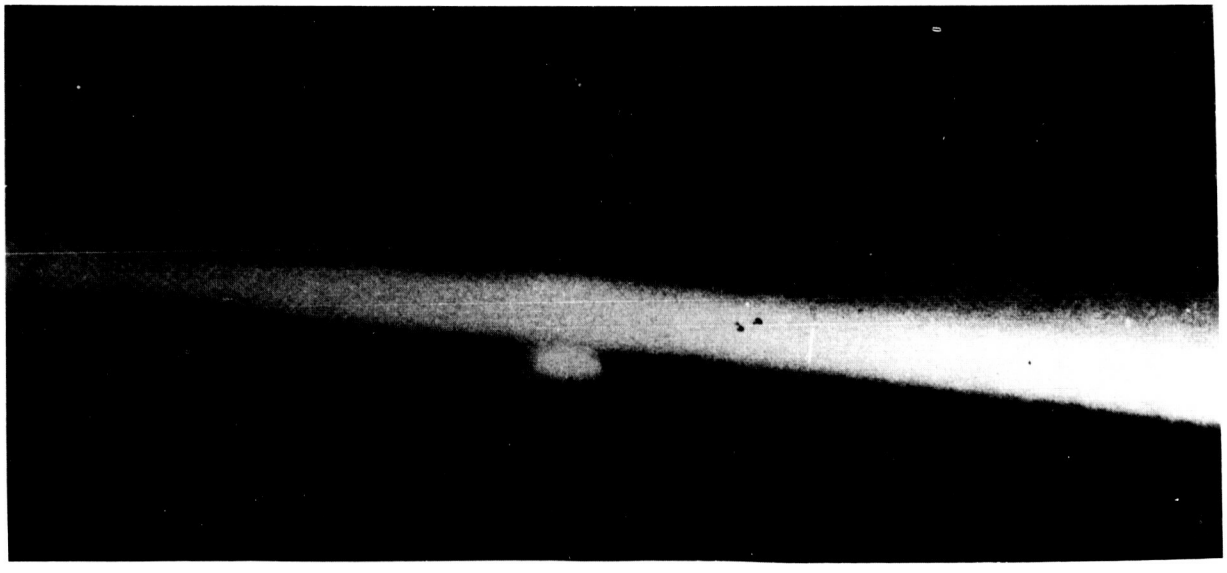


FIGURE 4.—Photograph of setting sun taken by Carpenter on the MA–7 orbital flight.

the capsule in orbit—has been demonstrated by direct observation on the MA–6 and MA–7 flights. However, it is hoped that on future missions photographs with precise observation times; and perhaps measurements of the apparent vertical and horizontal diameters with a sextant will be feasible.

ACKNOWLEDGMENTS

The authors take pleasure in acknowledging the assistance of Mr. James Donegan, Goddard Space Flight Center, for making available the final orbit elements and to Mr. Frederick B. Shaffer, Goddard Space Flight Center, for programing the orbit on the 7090 computer.

REFERENCES

1. GARFINKEL, B., "An Investigation in the Theory of the Astronomical Refraction," *Astronom. J.*, **50**(8): 169–179, February 1944.
2. The Rocket Panel, "Pressures, Densities, and Temperatures in the Upper Atmosphere," *Phys. Rev.*, **88**(5): 1027–1032, Dec. 1, 1952.
3. International Council of Scientific Unions, Committee on Space Research, Preparatory Group for an International Reference Atmosphere, "CIRA 1961: COSPAR International Reference Atmosphere, 1961; Report Accepted at the COSPAR Meeting in Florence, April 1961, comp. by H. Kallmann-Bijl, R. L. F. Boyd, et al., Amsterdam: North-Holland Publ. Co., 1961.
4. CHAUVENET, W., "A Manual of Spherical and Practical Astronomy," 5th ed., vol. 1, New York: Dover Publ., 1960.

THE EARLY EVOLUTION OF THE SUN

DILHAN EZER AND A. G. W. CAMERON

Goddard Space Flight Center

Hayashi has predicted that the early contracting sun should be highly luminous and fully convective, if one makes a correct choice of the photospheric boundary condition in constructing solar models in the contracting stages. We have constructed a sequence of models which fully confirm his predictions. The maximum radius of the protosun consistent with gravitational stability is $57R\odot$; this has a luminosity of about 450 times that of the sun. As the protosun shrinks its stays fully convective until a radius of less than $3R\odot$ is reached; the luminosity continues to decrease until a radius of $1.7R\odot$ is reached. The sun requires about 2 million years to contract onto the main sequence, but this number is very uncertain because of the unsatisfactory state of convection theories. If the primitive solar material has the terrestrial ratio of deuterium to hydrogen, then a further 3×10^5 years is required to burn the deuterium. These numbers are very much less than previous studies had indicated. The degree of lithium burning in the outer convection zone of the sun is highly uncertain owing also to the unsatisfactory state of convection theories.

INTRODUCTION

The early contracting phase of solar evolution is of great interest to any theory of formation of the solar system (see for example Hoyle, 1960; Cameron, 1962a, 1962b). There is a tendency on the part of scientists attempting to reconstruct the history of the planets and meteorites to assume that the sun had its present properties, or at least that its luminosity was not greatly different from the present one, during the earliest history of the solar system. We shall see in the course of this work that such an assumption is questionable.

The first detailed study of an evolutionary sequence of contracting solar models was carried out by Henyey, Lelevier, and Levee (1955). This work indicated that the sun gradually increased in luminosity as it contracted. When nuclear energy generation sets in and halts the contraction near the main sequence, a moderate decline in luminosity takes place. The solar models of Henyey et al. did not make provision for the possibility of an outer convection zone.

A more recent study has been carried out by Brownlee and Cox (1961). These authors improved the calculations of Henyey et al. by using improved opacities for the solar interior, by incorporating an outer convection zone in the models, and by putting in all the sources of nuclear energy generation. However, they used only a crude surface boundary condition. The evolutionary paths followed by their models have a general resemblance to those of Henyey et al.

Recently Hayashi (1962) has pointed out that the surface boundary conditions in the more extended models of Henyey et al. and of Brownlee and Cox are not correct. The density of matter in the photospheric layers of such models is much too small. In the photosphere of the sun the opacity of the material must be such that the photospheric layer lies at about one optical depth in the solar atmosphere. Thus Hayashi concluded that a readjustment of the contracting solar models would be necessary in which the density of the solar photospheric layer would be greatly increased.

Such a readjustment is possible only if there is a complete reorganization of the density-radius relation throughout the model. The models of Brownlee and Cox had rather shallow surface convection zones; an increase in surface density would require that the surface convection zone

should extend much farther into the interior, possibly all the way to the center. Since convection is an exceedingly efficient mechanism of energy transport, the luminosity of the model would thus depend entirely upon the rate at which energy could be radiated away from the surface. Now, the opacity of solar material increases as the temperature increases, at least in the range 3×10^3 to 10^4 °K. Thus a high surface temperature also assists the model to attain a satisfactory photospheric boundary condition. However, it should be noted that this leads to the expectation that the solar luminosity was much higher than at present during the contraction phase.

The present work was undertaken to investigate these predictions of Hayashi. The results reported here are preliminary, in the sense that isolated solar models have been calculated rather than evolutionary sequences of models. No nuclear energy generation has been included, so assumed to be in homologous contraction, so that the luminosity derives entirely from the release of gravitational potential energy. It was necessary to make an assumption about the distribution of this energy source. All models were that the relative structure would not change in an infinitesimal contraction. With assumptions of this sort it is easier to explore the consequences of variations in interesting physical parameters entering into the models.

THE BASIC EQUATIONS OF STELLAR STRUCTURE

Stellar models are calculated on the assumption that the star is spherically symmetric and in hydrostatic equilibrium. The equations that must be satisfied by this structure are:

$$dM(r)/dr = 4\pi r^2\rho \tag{1}$$

$$dP(r)/dr = -G[M(r)/r^2]\rho \tag{2}$$

$$dL(r)/dr = 4\pi r^2\rho\epsilon \tag{3}$$

For radiative equilibrium:

$$\frac{dT}{dr} = -\frac{3\chi\rho}{4ac}\frac{1}{T^3}\frac{L(r)}{4\pi r^2} \tag{4a}$$

For convective equilibrium:

$$\frac{1}{T}\frac{dT}{dr} = \frac{\Gamma-1}{\Gamma}\frac{1}{P}\frac{dP}{dr} \tag{4b}$$

In the above equations, P is the total pressure (gas plus radiation), T is the temperature, $M(r)$ is the mass within a sphere of radius r, $L(r)$ is energy crossing the surface of a sphere of radius r per second, ϵ is the energy produced per gram per second, χ is the absorption coefficient (cm^2/gm), and Γ is the effective ratio of the specific heats. It differs from $\gamma = c_p/c_v$ through the inclusion of the effects of dissociation and ionization of the gas, and it varies throughout the convection zone of the stars. The assumption $\Gamma = \text{constant} = 5/3$ corresponds to the assumption that the stellar material is wholly neutral or completely ionized.

Equations (1–4) are more easily integrated in electronic computers if they are expressed in logarithmic form. Neither the pressure nor the mass is very suitable to take as an independent variable all the way from surface to the center. Near the photosphere, where the change in pressure is rapid compared to the change in mass, and the energy transport by convection should be treated in the proper way, it is better to use pressure as an independent variable. Then equations (1–4) can be written in logarithmic form with pressure as an independent variable:

$$d\ln r/d\ln P = d\mathbf{r}/d\mathbf{P} = -\exp(\mathbf{r}+\mathbf{P}-\ln G-\mathbf{M}-\boldsymbol{\varrho})$$

$$d\ln M/d\ln P = d\mathbf{M}/d\mathbf{P} = -\exp[\ln(4\pi/G)+4\mathbf{r}+\mathbf{P}-2\mathbf{M}]$$

$$d\ln L/d\ln P = d\mathbf{L}/d\mathbf{P} = -\exp[\ln(4\pi/G)+4\mathbf{r}+\boldsymbol{\varepsilon}+\mathbf{P}-\mathbf{M}-\mathbf{L}]$$

Radiative temperature gradient:

$$d\ln T/d\ln P = d\mathbf{T}/d\mathbf{P} = \nabla_{\text{rad}} = \exp[\ln(3/16\pi GaC)+\ln\chi+\mathbf{L}+\mathbf{P}-\mathbf{M}-4\mathbf{T}]$$

The symbols indicate that $\mathbf{r} = \ln r$, $\mathbf{M} = \ln M$, etc. The temperature gradient in case of convection will be discussed later.

Even though the above equations can be used all the way from surface to the center, in cases where the change in mass becomes very large compared to the change in pressure, very small steps are needed. Then it is better to take mass as an independent variable. Our basic equations become:

$$d\mathbf{r}/d\mathbf{M} = \exp(\mathbf{M}-3\mathbf{r}-\boldsymbol{\varrho}-\ln 4\pi),$$

$$d\mathbf{P}/d\mathbf{M} = -\exp[\ln(6/4\pi)+2\mathbf{M}-\mathbf{P}-4\mathbf{r}],$$

$$d\mathbf{L}/d\mathbf{M} = \exp(\boldsymbol{\varepsilon}+\mathbf{M}-\mathbf{L}),$$

and

$$d\mathbf{T}/d\mathbf{M} = -\exp[\ln(3/64ac\pi^2)+\mathbf{L}+\mathbf{M} -4\mathbf{T}-4\mathbf{r}+\ln\chi],$$

provided

$$|\,d\mathbf{T}/d\mathbf{M}\,| < |\,[(\Gamma-1)/\Gamma]/(d\mathbf{P}/d\mathbf{M})\,|;$$

otherwise

$$d\mathbf{T}/d\mathbf{M} = [(\Gamma-1)/\Gamma]/(d\mathbf{P}/d\mathbf{M}).$$

It should be noted that the second form for the equation of energy transport is needed where the radiative temperature gradient exceeds the adiabatic temperature gradient. Because the logarithmic mass is used as an independent variable only deep in the stellar interior, we make here the assumption that the full energy flux is carried by convection and that the temperature gradient is the adiabatic one. This assumption is good in the deep interior but fails badly near the stellar surface where the logarithmic pressure is used as an independent variable.

THE GRAVITATIONAL ENERGY SOURCE

In this study of early solar evolution, the energy generation was assumed to be entirely due to release of the gravitational potential energy of the contracting solar mass.

The luminosity is

$$L = \int_0^R 4\pi r^2 \rho \epsilon(r)\, dr$$

where ϵ is energy released per gram of material per second due to contraction of the star.

$$\epsilon(r) = dL(r)/4\pi r^2 \rho\, dr$$

If U indicates the internal energy per gram of material and V the specific volume, then at each point in the star

$$\partial U/\partial t = -\epsilon(r) - P(\partial V/\partial t)$$

The change of internal energy is balanced by the energy loss and the work done by the pressure. The internal energy per gram of material at a point inside the star is

$$U = c_v T(r)$$

Assuming the ideal gas law governs the interior of the star,

$$U = [1/(\Gamma-1)]/[P(r)/\rho(r)],$$

where Γ is the ratio of specific heats at the point considered. If we replace V by its reciprocal ρ, then:

$$-\epsilon(r) = \frac{1}{(\Gamma-1)}\frac{\partial}{\partial t}\left(\frac{P}{\rho}\right) + P\frac{\partial}{\partial t}\left(\frac{1}{\rho}\right),$$

or

$$-\epsilon(r) = \frac{1}{(\Gamma-1)}\frac{1}{\rho}\frac{\partial P}{\partial t} - \frac{\Gamma}{(\Gamma-1)}\frac{P}{\rho^2}\frac{\partial \rho}{\partial t}. \quad (5)$$

If it is assumed that the star is contracting homologously, then the rate of change with time of pressure and density at every point is completely determined by the rate of change in the radius of the star in such a way that

$$(4P)^{-1}(\partial P/\partial t) = -R^{-1}(dR/dt)$$

and

$$(3\rho)^{-1}\partial\rho/\partial t) = -R^{-1}(dR/dt)$$

Inserting these values in Eq. (5):

$$\epsilon(r) = \frac{3\Gamma-4}{\Gamma-1}\frac{P(r)}{\rho(r)}\left(-\frac{1}{R}\frac{dR}{dt}\right).$$

Hence

$$L = 4\pi\left(-\frac{1}{R}\frac{dR}{dt}\right)\int_0^R \left(\frac{3\Gamma-4}{\Gamma-1}\right)P(r)r^2\, dr.$$

The rate of contraction may be written

$$-\frac{1}{R}\frac{dR}{dt} = \frac{L}{4\pi\int_0^R [(3\Gamma-4)/(\Gamma-1)]P(r)r^2\, dr}$$

$$= J\frac{LR}{GM^2}.$$

It may thus be seen that we are expressing the reciprocal time for the contraction rate as a fraction of the ratio of the luminosity to a measure of the gravitational potential energy. The quantity J may be called the contraction parameter; it is one of the fundamental parameters that had to be adjusted in order to obtain consistent solar models. It is by definition

$$J=\frac{GM^2}{4\pi R\int_0^R[(3\Gamma-4)/(\Gamma-1)]P(r)r^2\,dr}$$

It may easily be seen that the contraction parameter should not vary rapidly from one solar model to a later one. If the sun were a polytrope of index n with a constant ratio of specific heats throughout, then we would have

$$J=\frac{(\gamma-1)(5-n)}{3\gamma-4}.$$

OPACITIES

The computer programing code for stellar absorption coefficients and opacities (Cox, 1961), which was prepared at the Los Alamos Scientific Laboratory, was kindly made available for the present opacity calculations. The program was run on an IBM 7090 at the Institute for Space Studies for a temperature range of 2×10^3 °K through 5×10^7 °K and a density range of 10^{-12} through 10^3 gm/cm^3, with a mixture containing hydrogen, helium, and heavy elements having mass fractions 0.602, 0.376, and 0.022, respectively. The code requires use of the abundance of each element separately.

The code could be run for a mixture of not more than 11 elements. A modified version of an abundance table (Cameron, 1959) was adopted and the individual elements were combined to give the representative abundances shown in table I. The rather rare element K was retained because of its importance in supplying electrons for the formation of H^- at low temperatures.

The processes which contribute to the opacity calculation in the code are bound-bound absorption, bound-free absorption, free-free absorption, electron scattering, negative ion absorption, and electron conduction.

The bound-bound absorption by many lines close together (blanketing) has been important in stellar atmospheres and interiors. Its influence on the opacity depends on the position of the line in the emission spectrum and, to a lesser degree, the shape of the line. In stellar atmospheres Stark broadening of hydrogen lines is dominant, but in the deep interiors of the star collisional broadening is the most important below 10^6 °K; lines can increase the opacity by a factor of two or more. At the temperatures prevailing in the interiors of stars, the increase is about 10% of the total opacity.

The bound-free absorption depends on the equilibrium number of electrons which are bound in the various atomic states. When the ionization of one element is completed, no more bound-free absorption due to that element can occur. In the mixture under consideration, the bound-free absorption by the K shell electrons of iron keeps the opacity moderately large up to several million degrees.

At temperatures around 10^7 °K, most elements in astrophysical mixtures have been ionized, and free-free absorption in the field of hydrogen ions becomes very important. The nuclear charge is shielded by free electrons and ions resulting in a decrease in the free-free absorption. But these effects are about 10% or less at moderate densities and temperatures around 10^6 °K. When

TABLE I.—*Adopted Abundances of Elements* (*Silicon* $=10^6$)

Element	Abundance	Element	Abundance	Element	Abundance
H	3.2×10^{11}	O	2.9044×10^7	K	3.3×10^3
He	5.0×10^9	Ne	1.7×10^7	Ca	1.99×10^5
C	1.66×10^7	Al	9.5×10^4	Fe	2.629×10^5
N	3.0×10^6	Si	2.303×10^6		

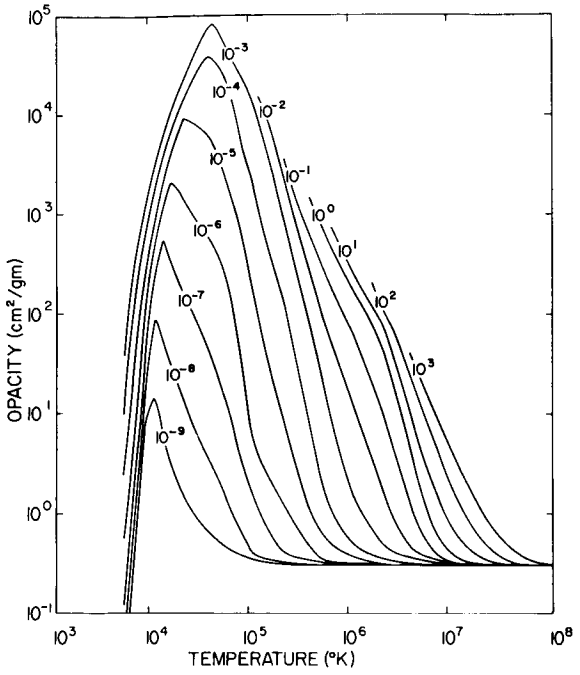


FIGURE 1.—Opacities for lines of indicated density (gm/cm³) calculated as a function of temperature from the Los Alamos Opacity Code.

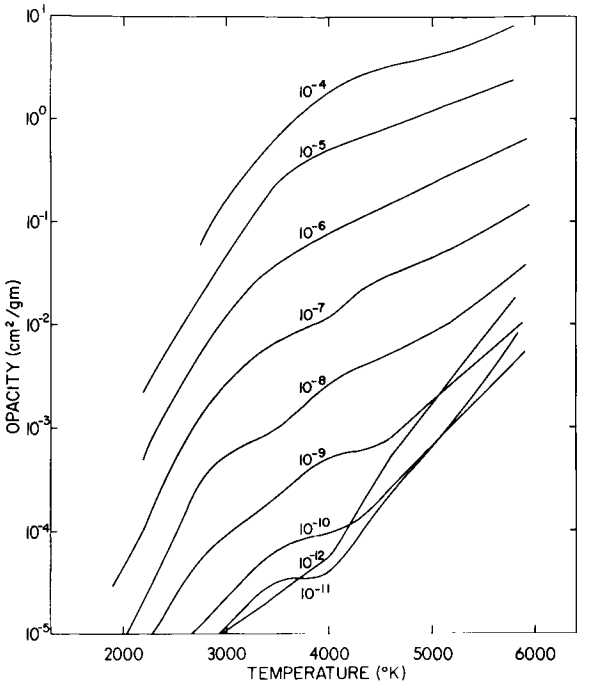


FIGURE 2.—Details of the opacities at low temperatures. The lines refer to the indicated densities.

electrons are degenerate, electron conduction also contributes to the transport of energy. But these cases are not important in the temperature and density ranges we are interested in.

The only negative ion of importance in astrophysics is that of hydrogen. At the low temperatures such as occur at the outer layer of stars, H^- ions formed from the capture of electrons from the metals keep the opacity still quite high. Both H^- bound-free and free-free absorptions contribute to the opacity. Actually H^- free-free absorption, which depends on the number of free electrons, is only important at wavelengths near or longer than the H^- bound-free edge.

Molecular absorption and Rayleigh scattering have not been included in the opacity calculation. Unfortunately very little is known about the absorption due to molecules that form in stellar atmospheres. Blanketing effects due to overlapping molecular bands should be very important at these low temperatures. Therefore, the opacities at low temperatures are very uncertain. Rayleigh scattering, which is included in Vitense's opacity calculations (Bohm-Vitense, 1951), is usually not very important. Only at very low temperatures and for low metal contents might it become important.

The calculated opacities require corrections due to the absorption by lines. A. N. Cox gives these corrections as a graph (private communication) and the application of these corrections increases the opacities by a factor of 1.1 to about 3.0 depending on the density and temperature combination.

Figure 1 shows the overall run of opacities calculated for different densities as a function of temperature. Figure 2 shows in more detail the opacities at low temperatures.

These opacities have been stored in the machine as a two-dimensional table corresponding to discrete values of $\ln T$ and $\ln\rho$. For a given $\ln T$ and $\ln\rho$ combination the corresponding opacity ($\ln\chi$) was obtained by linear interpolation.

EQUATION OF STATE

The total pressure is given by the sum of gas pressure plus radiation pressure,

$$P = P_g + P_r,$$

$$P_g = (1+\bar{x})\,(NkT/V)$$

$$P_r = \tfrac{1}{3}aT^3$$

where N is the number of atoms per unit volume, V is the specific volume, and a is the radiation constant; $\bar{x}$ is the mean degree ionization and dissociation defined by

$$\bar{x} = \sum \nu_i \eta_i - (x_4/2)\nu_1$$

where η_i are the degrees of ionization, x_4 is the fraction of hydrogen nuclei in molecular state, and ν_i are the abundances by number of various elements under consideration, with $\sum \nu_i = 1$. The partial electron pressure P_E and the partial pressure of atomic hydrogen P_H are given by

$$P_E = \bar{x}P_g/(1+\bar{x})$$

$$P_H = \frac{[1+\eta_1-(x_4/2)]\nu_1 P_g}{1+\bar{x}}$$

The density ρ is related to P and T through the equation of state

$$P_g = (\mathcal{R}/\mu)\rho T;$$

with the above definition for the mean molecular weight μ we have

$$\mu = \mu_0/(1+\bar{x})$$

with $\mu_0 = 4/(4X+Y)$, where X and Y are the mass fractions of hydrogen and helium, respectively.

As a chemical composition we use 0.602 for the mass fraction of hydrogen, 0.376 for helium, and 0.022 for the heavy elements. Hydrogen and helium have each been considered in three states: H_2, H, H^+ and He, He^+, He^{2+}. The dissociation equilibrium constant for molecular hydrogen is given by a polynomial of θ $(=5040/T)$ which is used in the calculations of Vardya (1961).

$$\log_{10}K(H_2) = \log(P^2_H/P_{H2} = 12.533505 - 4.92516440\theta + 0.056191273\theta^2 - 0.003268766 1\theta^3$$

Here P_H denotes the partial pressure of atomic hydrogen, and P_{H2} that of molecular hydrogen. The amount of molecular hydrogen is only calculated in the temperature range 1000 °K to 12009 °K.

The degree of ionization for each element is obtained from the Saha equation;

$$\frac{\eta_i}{1-\eta_i}\frac{\bar{x}}{1+\bar{x}}P_g = \text{Const}\,T^{5/2}\exp(-\chi_i/kT)$$

where χ_i is the ionization potential and k is the Boltzmann constant. At each integration step the program computes the degree of ionization and dissociation.

CONVECTION

The existence of an outer convection zone greatly affects the luminosity and radius of late type stars. These two quantities are the boundary conditions for the interior models of the stars. Therefore, convection should be treated as accurately as possible. We use Prandtl's mixing length theory as applied by E. Bohm-Vitense (1958).

In the ionization zone where the absorption coefficient increases rapidly and the adiabatic gradient decreases with increasing temperature, the temperature gradient $\nabla = d\ln T/d\ln P$ becomes very steep and the local adiabatic gradient ∇_{ad} falls below it. Consequently the gas will become unstable against convection. A temperature gradient will be set up in such a way that the total flux will be carried partly by radiation and partly by convection. Moreover, when convective instability occurs, the turbulent elements do not move strictly adiabatically. The rising and falling turbulent elements and their surroundings exchange heat; the temperatures of the elements differ from their surrounding material. If $\nabla' (= d\ln T'/d\ln P)$ shows the logarithmic temperature gradient of a moving turbulent element, then the main equations for the convective theory can be summarized as follows:

When the instability criterion

$$\nabla_{ad} < \nabla$$

is satisfied, we have

$$F_{rad} + F_{con} = F_{total} = L/4\pi R^2. \qquad (6)$$

The radiative flux is given by

$$F_{rad} = (16\sigma T^4/3\chi\rho H)\nabla. \qquad (7)$$

The convective flux is given by

$$F_{\text{conv}} = c_p \rho T \bar{v} (l/2H)(\nabla - \nabla'). \qquad (8)$$

Here, c_p is the specific heat per gram at constant pressure, $\bar{v}$ is the mean velocity of the rising or falling turbulent elements, $H = \Re T/\beta\mu g$ is the pressure scale height, and l is the mixing length, a distance which a moving element travels before disappearing into the surroundings. For most of this work we have assumed $l = H$. Also g is the gravitational acceleration and β is the ratio of the gas pressure to total pressure. The mean velocity is

$$\bar{v} = (l/2)[(g/H)C]^{1/2}(\nabla - \nabla')^{1/2}. \qquad (9)$$

Here C takes care of the change of the degree of ionization of matter with temperature and density changes and can be expressed by

$$C = 1 - [(\partial \ln \mu)/(\partial \ln T)].$$

The temperature of a moving element is given by a relation which represents the ratio of the excess energy content of the element to its radiation during its lifetime.

$$\frac{\nabla' - \nabla}{\nabla_{\text{ad}} - \nabla'} = c_p \rho T \frac{\chi \rho l}{24 \sigma T^4} \bar{v} \qquad (10)$$

The adiabatic temperature gradient for a material which is undergoing dissociation or ionization can be obtained from the entropy condition. The total internal energy of the volume V which contains matter and radiation is the sum of the kinetic and potential energies of the particles and the radiation energy of the volume:

$$U = [\tfrac{3}{2}(1+\bar{x})kT + \sum \nu_i \eta_i \chi_i] N + VaT^4.$$

For an adiabatic change,

$$dU + PdV = 0.$$

The above two equations, together with equation of state and the Saha equation for the ionization, make it possible to give a general expression for the adiabatic gradient in terms of known quantities

$$\nabla_{\text{ad}} = [(d \ln T)/(d \ln P)]_{\text{ad}} = F_1/F_2$$

$$F_1 = (1+\bar{x}) + (f_1/f_2) \sum \nu_i \eta_i (1-\eta_i) [\tfrac{5}{2} + (\chi_i/kT)]$$
$$+ 4[(1-\beta)/\beta][(1+\bar{x}) + (f_1/f_2) \sum \nu_i \eta_i (1-\eta_i)]$$

$$F_2 = \tfrac{5}{2}(1+\bar{x}) + \sum \nu_i \eta_i (1-\eta_i) [\tfrac{5}{2} + (\chi_i/kT)]^2$$
$$- \frac{[\sum \nu_i \eta_i (1-\eta_i)(\frac{5}{2} + \chi_i/kT)]^2}{f_2} + 4\frac{(1-\beta)}{\beta}$$
$$\times \left[4(1+\bar{x}) + \frac{f_1}{f_2} \sum \nu_i \eta_i (1-\eta_i) \left(\frac{5}{2} + \frac{\chi_i}{kT} \right) \right]$$

where $f_1 = \bar{x} + \bar{x}^2$ and $f_2 = 2\bar{x} + \bar{x}^2 - \sum \nu_i \eta_i^2$. They are taken as equal to $1+\bar{x}$ and $1+\bar{x} - (\sum \nu_i \eta_i / \mu_0)$ in the case of dissociation of hydrogen molecules under consideration. In this expression the variation with temperature of the molecular hydrogen partition function was neglected because of the small abundances of this molecule present in any of the models calculated in this paper. Should this assumption not be warranted, then one should use the thermodynamic functions given by Vardya (1961). The specific heat c_p, which is the partial derivative of U with respect to T, may be written as

$$c_p = (\Re/\mu_0) F_2.$$

The equations 6, 7, 8, 9, and 10 are the main equations which will give us the real temperature-pressure relation in the layers where convection sets in. These five equations determine the actual logarithmic temperature gradient ∇, the logarithmic temperature gradient of the turbulent element ∇', the velocity of the moving turbulent elements, the amount of flux carried by radiation F_{rad}, and that carried by convection F_{conv}, satisfying the condition that $F_{\text{rad}} + F_{\text{conv}}$ is equal to total flux. The convection will set in when $\nabla > \nabla_{\text{ad}}$. Since the radiative temperature gradient is the largest value that the actual temperature gradient can take, then ∇ must be between ∇_{rad} and ∇_{ad}. It is close to ∇_{rad} if the greater part of the energy is carried by radiation, and it has a value slightly above ∇_{ad} if most of the energy is carried by convection. On the other hand, if during the convection, the moving elements lost no energy, then the temperature gradient of the turbulent elements ∇' would be equal to ∇_{ad}. The moving elements never possess a temperature gradient steeper than the surrounding temperature gradient, since a rising element radiates energy to its neighborhood. Consequently the four temperature gradients should always maintain the relative

order

$$\nabla_{\rm rad} > \nabla > \nabla' > \nabla_{\rm ad}.$$

We use the following procedure in order to obtain the actual temperature gradient. Once T and P are known at a point, a subroutine calculates the degree of ionization, the mean molecular weight, the adiabatic gradient, the specific heat c_p, and C. The equations 6, 7, 8, 9, and 10 can be written in the following form:

$$w\nabla + Q(\nabla - \nabla')^{3/2} = F \tag{6'}$$

$$F_{\rm rad} = w\nabla \tag{7'}$$

$$F_{\rm conv} = Q(\nabla - \nabla')^{3/2} \tag{8'}$$

$$\bar{v} = q_1(\nabla - \nabla')^{1/2} \tag{9'}$$

$$(\nabla' - \nabla_{\rm ad}) = (9/4)(w/Q)(\nabla - \nabla')^{1/2} \tag{10'}$$

where

$$w = 16\sigma T^4/3\chi\rho H, \qquad q_1 = \tfrac{1}{2}l(g/H)^{1/2}C,$$

$$q_2 = c_p\rho T(l/2H), \qquad Q = q_1 q_2.$$

These equations can be combined to give a cubic equation

$$(Q/w)\Delta\nabla^3 + \Delta\nabla^2 + \tfrac{9}{4}(w/Q)\Delta\nabla + [\nabla_{\rm ad} - (F/w)] = 0 \tag{11}$$

in terms of $\Delta\nabla$, where $\Delta\nabla = (\nabla - \nabla')^{1/2}$.

If the radiative gradient becomes larger than the adiabatic gradient, the cubic equation is solved for $\Delta\nabla$ by Newton's method. Then equations 9′ and 10′ give us ∇ and ∇'. $F_{\rm rad}$ and $F_{\rm conv}$ are obtained with equations 7′ and 8′. We can check whether the condition (6) is satisfied. If it is not satisfactory, the values can be improved by iteration.

At great depths, the elements move almost adiabatically, and the actual temperature gradient approaches the adiabatic gradient. Convection is so efficient in transporting energy that we may use the convective temperature gradient as defined

$$d\mathbf{T}/d\mathbf{m} = [(\Gamma - 1)/\Gamma](d\mathbf{P}/d\mathbf{m})$$

provided that the actual values of Γ corresponding to the physical conditions existing at every point in the zone are taken into account.

ATMOSPHERIC CALCULATIONS

For an assumed mass, radius, and expected luminosity of the configuration, the effective temperature of the star,

$$T_{\rm eff} = L/4\pi R^2\sigma$$

is calculated. For the photospheric pressure:

$$dP_{\rm ph}/d\tau = g/\chi$$

where $g = GM/R^2$, χ is the absorption coefficient per gram of stellar material, and τ is the optical depth.

We start the calculations with an assumed photospheric density $\rho_{\rm ph}$, and $\chi(\rho_{\rm ph}, T_e)$ gives us the corresponding photospheric pressure at a chosen optical depth. Now the degree of ionization and dissociation allows us to define the mean molecular weight. From the equation of state:

$$P = (\Re/\mu)\rho T$$

a new ρ is determined. This procedure is repeated until $P_{\rm ph}$, $\rho_{\rm ph}$, T_e, $\chi(\rho_{\rm ph}, T_e)$, and τ are adjusted at a selected optical depth. From the study of the radiative transfer in stellar atmospheres, it is found that the effective temperature of the stars approximately corresponds to the temperature at the optical depth $\tau = 2/3$. We therefore used $\tau = 2/3$ in the calculation of our stellar models. In order to see the influence of different values of τ, we repeated the calculation for some models using $\tau = 1$ and $\tau = 0.5$. Due to the present uncertainties in the opacity at low temperatures, no attempt has been made to improve these crude atmospheric calculations.

INTEGRATION PROCEDURE

To get the final model for a star of fixed mass and radius, a computing program, using logarithmic variables, has been prepared for the 7090 computer which carries out the integrations and fitting procedure entirely automatically. With a fixed mass, radius, and trial luminosity of the star, atmospheric conditions provide us the boundary values for inward integrations. The luminosity L and the contraction rate J are the eigenvalues for these inward integrations. The final model cannot be obtained unless we have

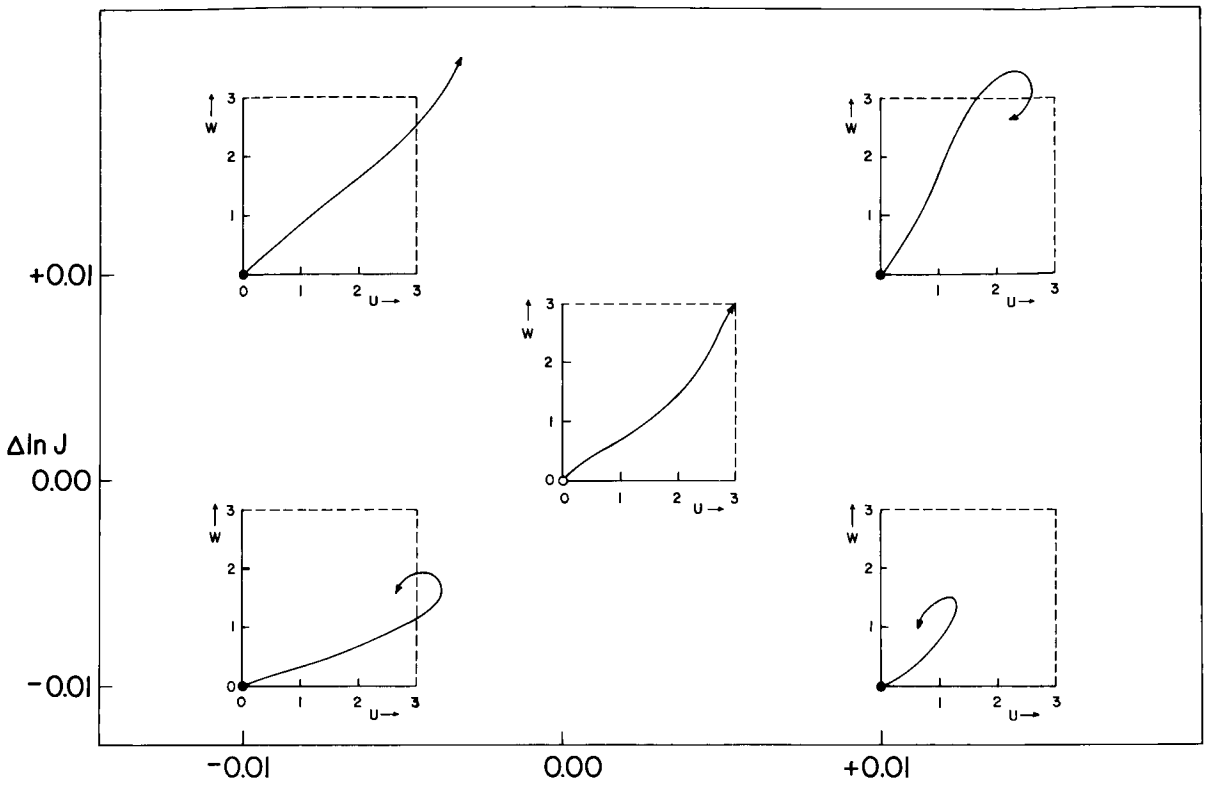


FIGURE 3.—Diagrams showing the behavior of the homologous variables U and W as a function of the assumed surface eigenvalues of the model, centered about the correct choice of these eigenvalues.

the right combination of L and J values which suits the internal structure of the star.

Integration starts from the surface using $\ln P$ as the independent variable. As a step size $\Delta \ln P = 0.1$ is stored in the machine, but the interval is halved or doubled according to the size of the derivatives. When the radiative temperature gradient gets larger than the simultaneously calculated adiabatic gradient, the actual temperature gradient is obtained as explained above. When $d\mathbf{M}/d\mathbf{P}$ is larger than a preassigned value, the program switches to the second set of equations in which the mass is used as an independent variable. Special care is taken in order to keep a uniform change in mass and pressure at the switchover point.

Preliminary trial inward integrations are carried out until the solutions reach within about 1% of the mass at the center. Stability of the integration is checked in terms of homology invariants

$$U = (d \ln M / d \ln r); \qquad W = (d \ln L / d \ln r)$$

which both should approach 3 near the center of the star. If the U value stays smaller than 3, the trial luminosity should be decreased; if it stays larger than 3 as the center is approached the trial luminosity should be increased. The systematic character of the W curve is also found to be a function of the trial eigenvalue of J. If W becomes smaller near the center, J should be increased; however, if it passes 3, the value should be decreased. Figure 3 schematically illustrates the character of these (U, W) curves as functions of the trial eigenvalues. The center sketch shows the behavior of the (U, W) curves with the right combination of eigenvalues L and J. The other figures illustrate the misbehavior of the curves with too large or too small combinations of the

trial eigenvalues. The program automatically checks U and W values and gives better trial values according to the mass point where the misbehavior starts and defines an upper and lower limit for each one as long as the other is behaving properly. The U value, which is the ratio of the local density to the mean density, needs special care in the ionization zones of hydrogen and helium. In these zones the local density might decrease with increasing geometrical depth, and our criteria for checking the stability of integrations by means of U might misdirect us; care is taken to avoid such misdirection.

The program tries to adjust the L and J values until the integrations are carried out to within 1% of the central mass. The program then obtains the trial center values for pressure P_c and temperature T_c by extrapolating from the values where the properly behaving inward integrations stop. The outward integration starts by series expansion at the mass $M_0=e^{-8}M$, where M is the total mass of the star, and is carried out up to the preassigned fitting point. The selection of right L and J values is very important since they affect the starting values of the outward integrations through the energy generation equation. If the trial eigenvalues of the integrations are far from the real conditions that the star should have, the fitting procedure is never achieved.

The fitting procedure is accomplished in the following manner: Two other inward integrations with small increments in L or J are performed, thus providing three inward integrations with the eigenvalues.

$$\begin{matrix} L, & J \\ L, & J+\delta J \\ L+\delta L, & J \end{matrix}$$

The comparison of these integrations at the preassigned fitting point gives the dependence of each dependent physical variables on the eigenvalues:

$$\delta Y_{i\ \mathrm{in}}/\delta L, \qquad \delta Y_{i\ \mathrm{in}}/\delta J$$

where Y_i $(i=1, \cdots, 4)$ are $\ln L$, $\ln R$, $\ln P$, and $\ln T$, respectively.

A similar set of four outward integrations, from the center to the fitting point are performed, with the trial eigenvalues

$$\begin{matrix} P_c+\delta P_c, & T_c, & L, & J \\ P_c, & T_c+\delta T_c, & L, & J \\ P_c, & T_c, & L+\delta L, & J \\ P_c, & T_c, & L, & J+\delta J. \end{matrix}$$

The results are compared with the outward integration of the eigenvalues P_c, T_c, L, J, in order to obtain the eigenvalue dependence

$$\delta Y_{i\ \mathrm{out}}/\delta P_c, \qquad \delta Y_{i\ \mathrm{out}}/\delta T_c,$$
$$\delta Y_{i\ \mathrm{out}}/\delta L, \qquad \delta Y_{i\ \mathrm{out}}/\delta J$$

of the physical variables at the fitting point.

The corrections ΔP_c, ΔT_c, ΔL, ΔJ to the trial eigenvalues of P_c, T_c, L, and J, which will give us smaller differences at the fitting point, can be obtained by solving four simultaneous linear equations,

$$\Delta D_i=\frac{\delta Y_{i\ \mathrm{out}}}{\delta P_c}\Delta P_c+\frac{\delta Y_{i\ \mathrm{out}}}{\delta T_c}\Delta T_c$$
$$+\left(\frac{\delta Y_{i\ \mathrm{out}}}{\delta L}-\frac{\delta Y_{i\ \mathrm{in}}}{\delta L}\right)\Delta L+\left(\frac{\delta Y_{i\ \mathrm{out}}}{\delta J}-\frac{\delta Y_{i\ \mathrm{in}}}{\delta J}\right)\Delta J$$

where ΔD_i $(i=1, \cdots, 4)$ are the differences at the fitting point between inward and outward integrations in physical variables. The procedure is repeated using better trial values, until the differences at the fitting point become smaller than a desired value.

RESULTS

Twenty-seven models for the early contracting sun, each corresponding to a radius expressed as a multiple of the present solar radius, have been constructed. The results of the computations have been summarized in table II for fully convective models and in table III for partially convective models. Each column shows, in turn, the radius in terms of solar radius, the luminosity in terms of solar luminosity, the effective temperature of the model, the contraction parameter, the central temperature of the model, the center and surface densities, and the ratio of the central density to the mean density of the model.

For all models above 2.5 solar radii, the convection is complete all the way to the center.

TABLE II.—*Characteristics of Fully Convective Models*

$R/R\odot$	$L/L\odot$	T_e (°K)	J	T_{center} (°K)	ρ_{center} (gm/cm³)	$\rho_{surface}$ (gm/cm³)	$\rho_c/\bar{\rho}$
3	3.470	4535	2.107	2.923×10^{6}	3.721×10^{-1}	1.829×10^{-7}	7.12
5	8.790	4432	2.080	1.779×10^{6}	8.301×10^{-2}	1.098×10^{-7}	7.35
10	28.677	4212	2.041	9.148×10^{5}	1.228×10^{-2}	6.197×10^{-8}	8.70
20	88.910	3941	2.015	4.758×10^{5}	1.606×10^{-3}	3.638×10^{-8}	9.10
25	124.68	3847	2.020	3.870×10^{5}	8.794×10^{-4}	3.089×10^{-8}	9.74
30	165.30	3768	2.062	3.269×10^{5}	5.377×10^{-4}	2.722×10^{-8}	10.3
50	343.91	3505	2.607	1.902×10^{5}	1.102×10^{-4}	2.048×10^{-8}	9.76
55	414.50	3502	2.436	1.848×10^{5}	1.059×10^{-4}	1.855×10^{-8}	12.48
60	475.13	3469	2.824	1.622×10^{5}	7.194×10^{-5}	1.782×10^{-8}	11.02
75	651.82	3358	3.024	1.389×10^{5}	4.932×10^{-5}	1.558×10^{-8}	14.74
100	976.72	3218	3.699	9.609×10^{4}	1.872×10^{-5}	1.313×10^{-8}	13.27
200	2555.0	2894	3.156	4.723×10^{4}	1.637×10^{-6}	8.595×10^{-9}	9.28
400	7325.4	2663	2.785	1.990×10^{4}	1.341×10^{-7}	7.374×10^{-9}	6.08
450	8107.9	2575	2.486	1.722×10^{4}	6.905×10^{-8}	8.464×10^{-9}	4.20
500	10730	2621	2.604	1.556×10^{4}	3.964×10^{-8}	6.444×10^{-9}	2.84
600	16078	2646	3.329	1.348×10^{4}	1.448×10^{-8}	4.798×10^{-9}	0.53
750	27900	2716	3.165	1.166×10^{4}	3.875×10^{-9}	2.997×10^{-9}	0.27
1000	62160	2874	0.595	1.101×10^{4}	1.027×10^{-9}	1.445×10^{-9}	0.73

When the sun contracts to a radius smaller than three solar radii, the radiative core starts to develop. Additional characteristics for these models are indicated in table IV; the second and third columns give the temperature and density at the bottom of the convection zone, the fourth column the mass fraction inside the radiative core which increases with decreasing radius; and the fifth column is the fraction of the radius covered by the radiative core.

In table V, we show the gravitational potential energy, the thermal energy, and ionization and dissociation energy, in ergs, for the indicated models. Meanwhile, the fractional mass in the form of molecules, neutral and ionized hydrogen, and neutral, singly, and doubly ionized helium are given. For the models $R/R\odot > 57$, the absolute value of the gravitational potential energy is less than the sum of the thermal, ionization, and dissociation energies of the material.

These models are in the region of instability against gravitational collapse (Cameron, 1962a).

TABLE III.—*Characteristics of Partially Convective Models*

$R/R\odot$	$L/L\odot$	T_e (°K)	J	T_{center} (°K)	ρ_{center} (gm/cm³)	$\rho_{surface}$ (gm/cm³)	$\rho_c/\bar{\rho}$
1.00	4.06	8169	0.748	2.270×10^{7}	479	1.293×10^{-8}	339
1.20	3.73	7306	0.779	1.888×10^{7}	263	3.035×10^{-8}	322
1.50	1.70	5374	1.234	9.622×10^{6}	41.62	2.171×10^{-7}	99.5
1.60	1.09	4658	2.096	5.220×10^{6}	4.644	3.456×10^{-7}	13.4
1.75	1.15	4507	2.080	4.574×10^{6}	2.210	3.545×10^{-7}	8.39
1.85	1.32	4538	2.088	4.453×10^{6}	1.796	3.228×10^{-7}	8.06
2.00	1.56	4550	2.120	4.206×10^{6}	1.302	2.915×10^{-7}	7.39
2.25	1.98	4551	2.110	3.848×10^{6}	0.884	2.533×10^{-7}	7.13
2.50	2.44	4548	2.095	3.512×10^{6}	0.647	2.242×10^{-7}	7.16

TABLE IV.—*Some Physical Characteristics at Base of Convection Zones*

$R/R\odot$	T (°K)	ρ	(M_r/M)	(r/R)
1.00	6.68×10^{4}	2.66×10^{-8}	1.00	1.00
1.20	7.10×10^{4}	5.08×10^{-7}	.99	.99
1.50	1.46×10^{5}	7.50×10^{-2}	.97	.70
1.60	2.58×10^{6}	7.68×10^{-1}	.60	.51
1.75	3.00×10^{6}	1.15	.52	.44
1.85	3.29×10^{6}	0.96	.40	.40
2.00	3.31×10^{6}	0.94	.15	.33
2.25	3.35×10^{6}	0.78	.08	.17
2.50	3.38×10^{6}	0.64	.002	.05

Hence, they are only of formal interest. The collapse of the protosun continues until the hydrogen is fully ionized, 32% of helium is singly ionized, and 66% of helium is doubly ionized. This corresponds to a radius of $57R\odot$, at which point gravitational stability sets in. The further contraction of the protosun occurs on the Kelvin-Helmholtz time scale.

In figure 4, the position of the models in the Hertzsprung-Russell diagram is shown. The threshold for stability corresponds to $R=57R\odot$. The luminosity of the protosun, corresponding to this radius, is about $450L\odot$. For all models above 2.5 solar radii, the convection is complete all the way to the center. When the protosun contracts to 3 solar radii, its central temperature reaches about 2.9×10^{6} °K and the luminosity is about 3.47 times solar luminosity. With further contraction a radiative core starts to develop at the center. When the sun contracts to 2 solar radii, the radiative core covers about 15% of the mass and 33% of the radius. The luminosity reaches its minimum value when the radius becomes about 1.7 times the solar radius. At that point the radiative core extends to about 55% of the mass.

With further contraction, the luminosity increases slightly, and when the radius has contracted to 1.5 solar radii, the radiative core covers about 97% of the mass and 70% of the radius. At that point, the central temperature reaches about 9.6 million degrees. The last part of the

TABLE V.—*Details of Internal Energies and States of Ionization*

$R/R\odot$	E_{grav}	$E_{thermal}$	$E_{ionization}$	$X(H_2)$	$X(H)$	$X(H^+)$	$Y(He)$	$Y(He^+)$	$Y(He^{2+})$
1	9.941×10^{48}	5.150×10^{48}	2.957×10^{46}	----------	neg	0.602	neg	neg	0.376
2	1.714×10^{47}	8.636×10^{47}	2.941×10^{46}	----------	0.002	.600	neg	0.003	.373
2.25	1.527×10^{48}	7.653×10^{47}	2.938×10^{46}	----------	.002	.600	neg	.003	.373
2.5	1.318×10^{48}	6.545×10^{47}	2.938×10^{46}	----------	.002	.600	neg	.004	.373
3.0	1.150×10^{48}	5.772×10^{47}	2.936×10^{46}	----------	.002	.600	neg	.004	.373
5	6.991×10^{47}	3.529×10^{47}	2.932×10^{46}	----------	.002	.600	neg	.006	.370
10	3.492×10^{47}	1.799×10^{47}	2.818×10^{46}	----------	.002	.600	neg	.014	.362
20	1.868×10^{47}	9.723×10^{46}	2.874×10^{46}	----------	.002	.600	neg	.029	.347
25	1.522×10^{47}	8.024×10^{46}	2.842×10^{46}	----------	.002	.600	neg	.042	.334
30	1.287×10^{47}	6.926×10^{46}	2.809×10^{46}	----------	.002	.600	0.001	.053	.322
50	7.501×10^{46}	4.738×10^{46}	2.669×10^{46}	----------	.002	.600	.004	.107	.265
55	7.311×10^{46}	4.611×10^{46}	2.631×10^{46}	----------	.002	.600	.006	.111	.259
60	6.426×10^{46}	4.298×10^{46}	2.583×10^{46}	----------	.002	.600	.008	.131	.237
75	5.550×10^{46}	3.9601×10^{46}	2.477×10^{46}	----------	.003	.599	.018	.155	.203
100	3.940×10^{46}	3.454×10^{46}	2.263×10^{46}	----------	.003	.599	.027	.227	.122
200	1.788×10^{46}	1.450×10^{46}	1.718×10^{46}	1.6×10^{-5}	.037	.565	.153	.222	.001
400	7.487×10^{45}	1.062×10^{46}	1.067×10^{46}	3.2×10^{-4}	.191	.411	.368	.008	neg
450	6.244×10^{45}	9.581×10^{45}	9.394×10^{45}	5.1×10^{-4}	.240	.362	.375	.001	neg
500	5.375×10^{45}	8.063×10^{45}	8.161×10^{45}	6.5×10^{-4}	.238	.364	.376	neg	neg
600	4.138×10^{45}	5.657×10^{45}	6.364×10^{45}	1.3×10^{-3}	.3542	.247	.376	neg	neg
750	2.892×10^{45}	4.088×10^{45}	5.193×10^{45}	4.4×10^{-3}	.4005	.2018	.376	neg	neg
1000	2.188×10^{45}	8.181×10^{45}	3.108×10^{45}	6.8×10^{-3}	.476	.119	.376	neg	neg

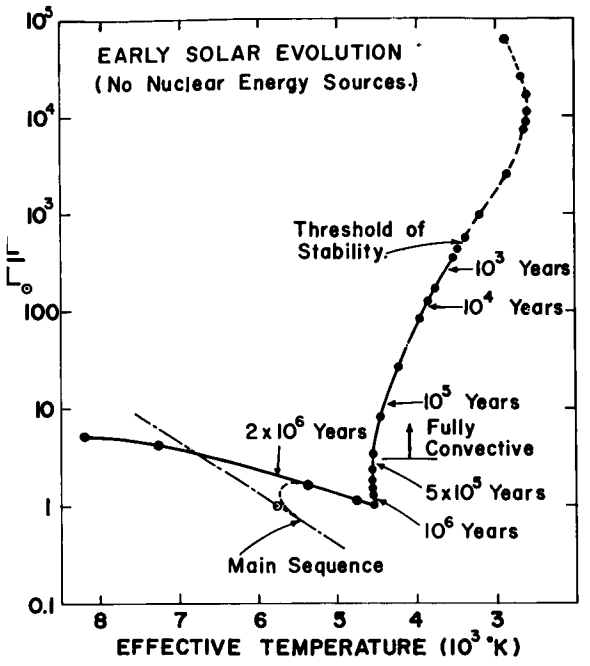


FIGURE 4.—The Hertzsprung–Russell diagram for the models of the contracting sun calculated in this paper, with the mixing length equal to the pressure scale height. Above the indicated threshold of stability the models are of academic interest only since the thermal, ionization, and dissociation energies exceed the magnitude of the gravitational potential energy. The models are fully convective above the line of evolution away from the threshold of stability as indicated at various points along the track. The main sequence and the present position of the sun are indicated by the dash-dot line on the lower left and by the circle. The sun probably approaches the main sequence approximately along the dashed line shown beyond the third last calculated model.

calculated track, which corresponds to the points $1.2R\odot$ and $1R\odot$, has no physical significance since we did not consider the nuclear energy generation in our model calculations. By the time the sun contracts to that extent the central temperature has become high enough to take into account the energy generation by nuclear sources. The start of nuclear energy generation will change the route of the track and should bring the sun down into the initial main sequence, probably about as indicated by the dashed portion on the lower left of figure 4.

In our calculations, we assumed that the effective temperature of the star corresponds to the actual temperature at an optical depth 2/3. It can be argued that this is not a good assumption. In order to test the sensitivity of this choice, we reconstructed some of the models by taking this value as 0.5 and 1. The results for a partially convective model $2R\odot$ and a fully convective model $20R\odot$ corresponding to the optical depths $\tau=0.5$, 2/3, and 1 are summarized in table VI. An inspection of these results indicates that the differences due to this choice can be considered as insignificant.

Prandtl's mixing-length theory in the form given by E. Vitense (1958) has been used in treating the convection in spite of some severe defects in representing the convective heat fluxes near the surface layers (Spiegel, 1962; Simoda, 1961). The mixing length used here is the vertical distance that a moving element travels before it dissolves. It is usually taken as a constant multiple of the pressure scale height. There are still great uncertainties about the appropriate choice of this constant. Therefore, it has been customary to choose this ratio as a disposable parameter. We constructed most of our models by taking this parameter as unity. In order to see the influence of different values for the ratio of the mixing length to the pressure scale height, $\alpha=l/H$; we constructed models $2R\odot$, also using the values

TABLE VI.—*Models with Effective Temperature at Different Optical Depths*

$R/R\odot$	ρ	T_e	$\rho_{surface}$	L	J	T_c	ρ_c
2	0.5000	4528	2.507×10^{-7}	5.792×10^{33}	2.118	4.203×10^{6}	1.314
2	0.6666	4550	2.915×10^{-7}	5.903×10^{33}	2.120	4.206×10^{6}	1.303
2	1.0000	4596	3.556×10^{-7}	6.145×10^{33}	2.118	4.218×10^{6}	1.303
20	0.5000	3926	3.120×10^{-8}	3.272×10^{35}	2.020	4.760×10^{5}	1.625×10^{-3}
20	0.6666	3941	3.638×10^{-8}	3.323×10^{35}	2.015	4.758×10^{5}	1.606×10^{-3}
20	1.0000	3971	4.486×10^{-8}	3.425×10^{35}	2.022	4.754×10^{5}	1.619×10^{-3}

TABLE VII.—*Models with Different Mixing Length Assumptions*

$R/R\odot$	l/H	T_e	$\rho_{surface}$	L	T_c	ρ	T (conv)	(r/R) (conv)	(M_r/M) (conv)	$L/L\odot$
2	0.5	4091	4.397×10^{-7}	3.858×10^{33}	4.140×10^{6}	1.452	3.02×10^{6}	0.42	0.25	1.02
2	1	4550	2.915×10^{-7}	5.903×10^{33}	4.206×10^{6}	1.303	3.31×10^{6}	0.33	0.15	1.56
2	2	5119	1.915×10^{-7}	9.450×10^{33}	4.347×10^{6}	1.239	3.78×10^{6}	0.23	0.07	2.50
2	3	5381	1.555×10^{-7}	1.156×10^{34}	4.331×10^{6}	1.239	4.11×10^{6}	0.11	0.01	3.06
20	0.5	3355	6.722×10^{-8}	1.745×10^{35}	4.718×10^{5}	1.578×10^{-3}	----------	------	--------	46.164
20	1	3941	3.638×10^{-8}	3.323×10^{35}	4.758×10^{5}	1.606×10^{-3}	----------	------	--------	87.91
20	2	4583	1.990×10^{-8}	6.077×10^{35}	4.790×10^{5}	1.659×10^{-3}	----------	------	--------	160.77
20	3	4970	1.480×10^{-8}	8.410×10^{35}	4.807×10^{5}	1.668×10^{-3}	----------	------	--------	222.48

0.5, 2, and 3 for α. The results are summarized in table VII. Figures 5 and 6 show the change of luminosity and effective temperature. The increase in α results in an increase in the luminosity of the star. The relative increase in luminosity is slightly greater in $20R\odot$ which is a fully convective model. These results have special importance in the sense that the luminosity of protosun might be higher than the calculated one, if for α a more realistic value turns out to be larger than one.

With a larger ratio of the mixing length to the pressure scale height causing an increase in luminosity, also there is an accompanying extension of the convection zone towards the center of the star. In accordance with this, the temperature at the bottom of the convective zone increases, and reaches about 4.1×10^{6} °K, for the model $2R\odot$. The radiative core covers only 1% of the mass with the value $\alpha=3$ while it is about 15% of the mass with $\alpha=1$. These results are shown in figure 7.

Convection sets in at the optical depths of $\tau=3.19$, 3.25, 2.05, and 1.94 for $\tau=0.5$, 1, 2, and 3, respectively for $2R\odot$.

In the transition regions where the density and the specific heat are low, the superadiative gradient ($\nabla-\nabla'$) and convective velocity $\bar{v}$ show a sudden increase in order to carry the required amount of flux. In figure 8 we have plotted the average convective velocity $\bar{v}$ versus $\log P$ for $20R\odot$. The increase in the ratio of mixing length to pressure scale height results in higher convective velocities; the convective velocity approaches that of sound and might initiate shock-wave phenomena if a large value of the ratio α should be the correct one.

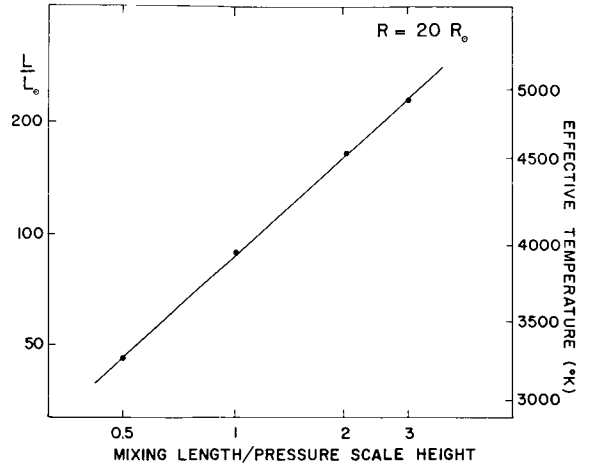


FIGURE 5.—The relation between luminosity and mixing length for a solar model with 20 times the present radius.

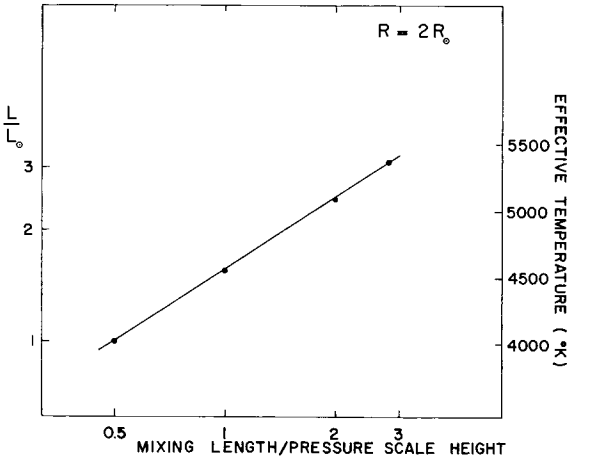


FIGURE 6.—The relation between luminosity and mixing length for a solar model with twice the present radius.

We have also investigated the effect of using the density scale height as a measure of the mixing length. The procedure here is somewhat complicated by the fact that the models so far described have a density inversion immediately below the transition region in the photosphere where convection begins. This arises from the effects of an increasing opacity with depth and the corresponding steepening of the temperature gradient necessary to carry the radiative flux in the region before convection becomes efficient. To avoid these complications we took as the mixing length in the

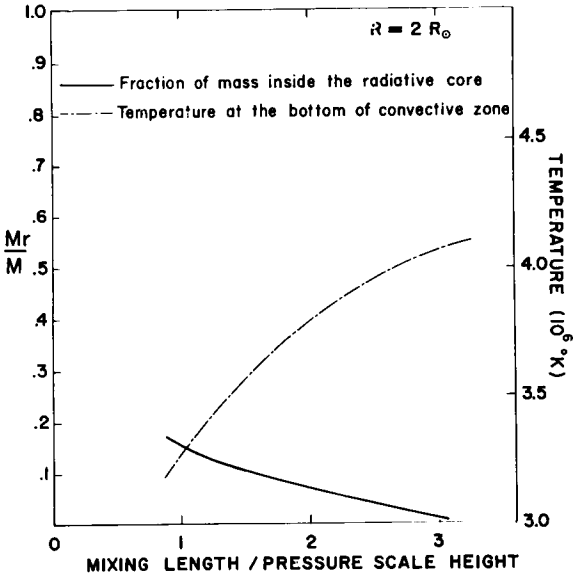


FIGURE 7.—Densities at the bottom of the outer convective zone in the solar model with twice the present radius, as a function of the assumed mixing length.

transition zone the actual distance required for the density to increase by a factor e below the point where convection commences. At greater depths the local density scale height was used.

With these changes the density inversion nearly disappeared. A model with radius $2R\odot$ had a luminosity corresponding to a choice $\alpha=1.4$ for the pressure scale height. A model with radius $20R\odot$ had a luminosity corresponding to a choice $\alpha=1.8$. These results are consistent with some calculations for a solar model carried out by Norman Baker (private communication).

The contraction time t for contraction between

$$t=\int_{R_1}^{R_2}\frac{d[E_{\mathrm{Gra}}-(E_{\mathrm{Ther}}+E_{\mathrm{ion}}+E_{\mathrm{disso}})]\,dR}{L\,dR}.$$

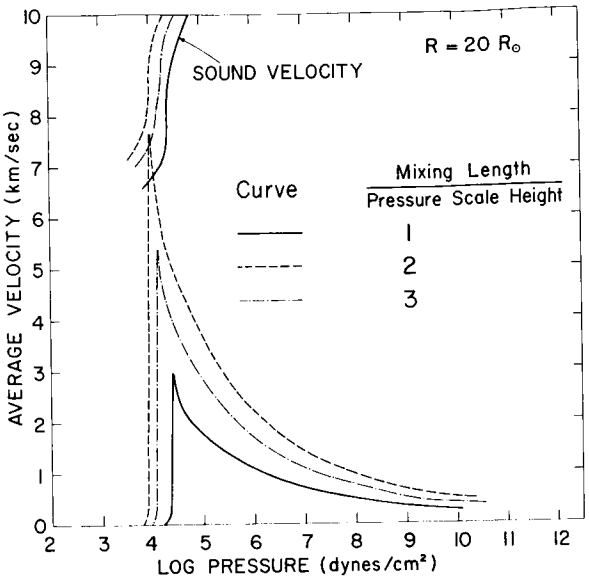


FIGURE 8.—The average velocity of the turbulent elements in the outer convective zone for a solar model of 20 times the present radius, shown for three assumptions concerning the mixing length. Also shown are the sound velocities at various points of the transition layers in the different models. The peak in the velocity curve is not realistically determined because the peak structure is narrow compared to the mixing length.

This integration has been carried out graphically. The luminosity and radius versus time have been plotted in figure 9, and also indicated in figure 4. It can be seen that the sun evolves very rapidly away from the threshold of stability. When it ceases to become fully convective its

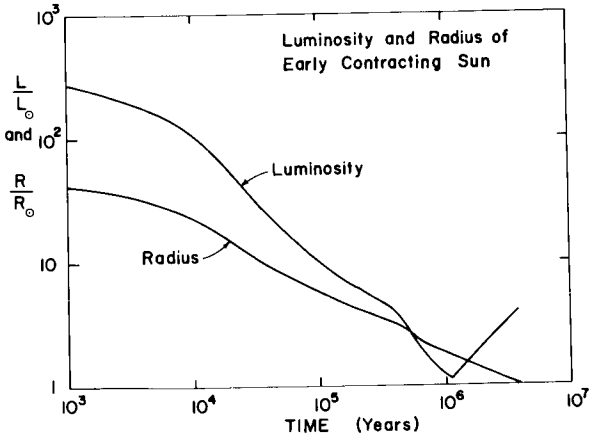


FIGURE 9.—The behavior of the luminosity and radius of the sun as a function of time beyond the threshold of stability.

age is about half a million years; the time required to contract from $57R\odot$ to $1.7R\odot$, the radius corresponding about the minimum luminosity, is about 1 million years. The sun reaches approximately its present position near the main sequence in about 2×10^6 years.

These times are much shorter than the earlier estimates of the age of the sun at the end of its stage of Helmholz–Kelvin contraction.

In the calculation of these results, no energy generation due to the deuterium burning has been included. Let us consider when the deuterium burning would occur as the primitive sun contracts.

The energy generation rate for deuterium burning is given by

$$\epsilon = 4.3\times10^{23}(x_H x_D \rho / T_6^{2/3}) \exp(-37.2/T_6^{1/3})$$

erg/gm sec

where x_H and x_D are the fractional hydrogen and deuterium abundances by mass, and T_6 is the temperature in unit of 10^6 °K. If the primitive sun had the terrestrial ratio of deuterium to hydrogen, this deuterium would be destroyed by thermonuclear reactions when the central temperature reaches about 800,000°K. When the sun contracts to 10 solar radii, the energy generation by deuterium burning is equal to 1.03×10^{35} erg/sec. At this radius its luminosity is 28 times its present value. The contraction would halt at this point while the deuterium is destroyed. Since the sun is fully convective, the material would be well mixed inside the sun and the destruction of deuterium would go on throughout the sun. This deuterium burning stage of the early solar evolution would add about 3×10^5 years to the contraction age indicated in figure 9.

We should also give some consideration to the present abundance of lithium in the sun. This abundance is much smaller, relative to that of similar elements such as sodium or potassium, than is found for the earth and meteorites. Thermonuclear reactions with hydrogen above 4,000,000°K rapidly destroy lithium. But we are determining the abundance of lithium in its surface layers. In order to find whether lithium in the surface layers should have been depleted by thermonuclear reactions, we must find the highest temperature which the convection zone attains

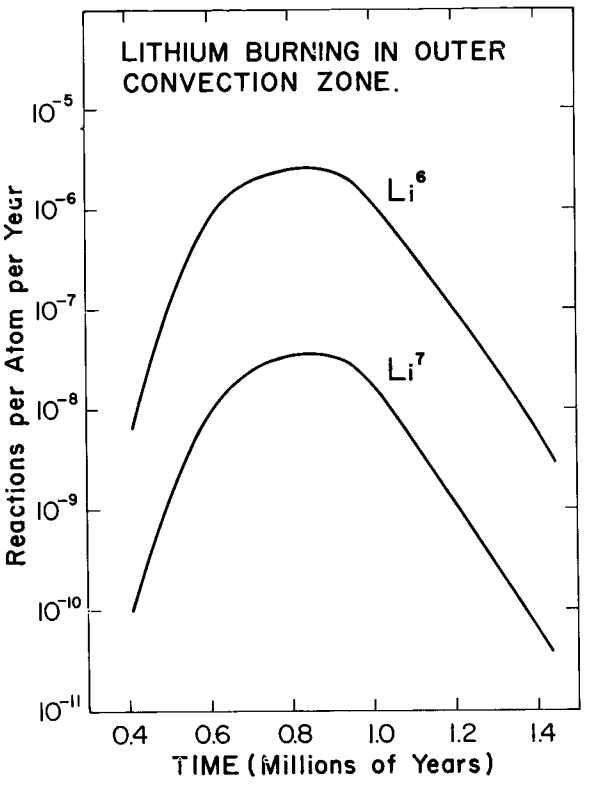


FIGURE 10.—The rate of lithium burning averaged throughout the outer convective zone for both isotopes of lithium, as a function of the evolutionary time of the solar models calculated with the mixing length assumed equal to the pressure scale height. The total depletion of the isotopes is obtained by integrating under these curves.

at the bottom so that the material can be subjected to these temperatures by convective motion.

We can estimate how much lithium would have been destroyed during the contraction phase of the sun. Lithium contains two stable isotopes, Li^6 and Li^7. The reaction rate for Li^6 (p, α) He^3 reaction per nucleus per second is

$$P/\rho x_H = [(5.96\times10^{12})/T_6^{2/3}] \exp(-84.149/T_6^{1/3})$$

The similar reaction rate for Li^7 (p, α) He^4 is

$$P/\rho x_H = [(1.20\times10^{11})/T_6^{2/3}] \exp(-84.731/T_6^{1/3})$$

If we assume that the mixing is sufficiently rapid to keep the material homogeneous in the convective zone, then the relative number of lithium nuclei which have been destroyed in the convective zone can be obtained by integrating

the reaction rates throughout the convection zone and dividing by the mass remaining inside the convective zone. For the fully convective model the surface material is well mixed with the center material, but the central temperatures have not risen above the critical value for the lithium thermonuclear reactions. When the sun contracts to $3R\odot$, the central temperature is high enough to consider the thermonuclear reactions with hydrogen. With further contraction the temperature at the bottom of the outer convective zone reaches the values listed in table IV. These temperatures are sufficiently high to deplete the Li^6 by a substantial factor (3.24), but the depletion of Li^7 is insignificant (1.4%). In figure 10 is shown the rate of lithium burning versus time.

In figure 7, we showed the change of temperature at the bottom of the convective zone with different assumed values of the ratio of the mixing length to the pressure scale height. With a larger ratio of the mixing length to the pressure scale height the temperature at the bottom of the convective zone increases. Consequently the depletion of lithium would be more rapid. With $\alpha=3$ for the $2R\odot$ model, the outer convective zone reaches a temperature well over 4×10^6 °K. At these high temperatures the depletion of Li^7 would be appreciable (factor 2.2). If the primitive sun contains about the same amount of lithium, relative to sodium and potassium, as do the earth and the meteorites, then we could obtain about the observed amount of depletion of lithium in the surface layers of the sun by assuming a fairly large value for the ratio of mixing length of pressure scale height in constructing models of the contracting phases.

Some of these results are quite similar to those independently obtained by R. Weymann and E. Moore (to be published). The principal difference is that we find a greater probability of lithium burning than they did, probably because our interior opacities are greater than the ones they used, and hence our outer convection zones are somewhat deeper.

CONCLUSIONS

These calculations confirm Hayashi's prediction that the early contracting sun should be highly luminous and fully convective. We find that the sun contracts to approximately its main sequence position in 2 million years, but this figure may be still further reduced owing to the great uncertainties in the solar lunimosity that we found associated with the theory of convection used. The maximum radius of the sun is $57R\odot$; models with larger radii are unstable against gravitational collapse. If the terrestrial ratio of deuterium to hydrogen also existed in the early protosun, then deuterium burning adds 3×10^5 years to the solar contraction time. We find that the outer convection zone of the sun very probably became hot enough to cause a significant depletion of Li^6, but the depletion of Li^7 is highly uncertain and depends on the details of the convection theory. These conclusions have considerable relevance for theories of the origin of the solar system, since it appears that planets forming in a solar nebula formed with the sun must be subjected to a high temperature environment.

ACKNOWLEDGMENTS

The authors are grateful to Dr. A. N. Cox for providing the Los Alamos opacity code and for consulting in regard to its use. They also wish to thank Mrs. Marilyn Golub and Mr. Leon Lefton for great assistance in carrying out the calculations that were performed on the Institute for Space Studies 7090 computer. One of us (D. Ezer) wishes to express appreciation to Dr. Robert Jastrow for the hospitality of the Institute for Space Studies during her tenure of a National Academy of Sciences-National Research Council Resident Research Associateship with the National Aeronautics and Space Administration.

REFERENCES

BROWNLEE, R. R., and COX, A. N., *Sky and Tel.*, **22,** 252 (1961).
CAMERON, A. G. W., *Astrophys. J.*, **129,** 676 (1959).
———, *Icarus*, **1,** 13 (1962*a*).
———, *Sky and Tel.*, **23,** 244 (1962*b*).
COX, A. N., Preprint (1961).
HAYASHI, C., *Publs. Astron. Soc. Japan*, **13,** 450 (1961).
HENYEY, L. G., LE LEVIER, R., and LEVEE, R. D., *Publs. Astron. Soc. Pacific*, **67,** 154 (1955).
HOYLE, F., *Quart. J. Roy. Astron. Soc.*, **1,** 28 (1960).
———, and SCHWARZSCHILD, M., *Astrophys. J. Suppl.*, **2,** 1 (1955).
SIMODA, M., *Publs. Astron. Soc. Japan*, **13,** 424 (1961).
SPIEGEL, E., To be published (1962).
VARDYA, M. S., *Astrophys. J.*, **133,** 107 (1961).
VITENSE, E. B., *Z. Astrophys.*, **28,** 81 (1951).
———, *Z. Astrophys.*, **32,** 135 (1953).
———, *Z. Astrophys.*, **46,** 108 (1958).

THE ROLE OF PLASMA INSTABILITIES IN THE ORIGIN OF SOLAR FLARES

R. K. JAGGI

Goddard Space Flight Center

We have investigated the role of plasma instabilities in the origin of a solar flare. Three types of stability problems are discussed: the magnetohydrodynamic instability of an infinitely conducting fluid, the instability due to the magnetic field gradients, and the finite conductivity instability. These instabilities are examined in a situation that occurs when two plasmas trapped in a pair of sunspot magnetic fields approach each other. The assumption of infinite conductivity gives rise to stable situations. The instability growth time in the magnetic field gradients that exist near sunspots is extremely large. The growth time for finite conductivity instability is of the order of a few seconds or minutes, and we therefore believe that the flare can be produced by finite conductivity instabilities in the solar atmosphere above a group of sunspots. Our analysis does not help explain the acceleration of charged particles to high energies; for this, the problem of acceleration must be considered separately.

1. INTRODUCTION

It has been suggested (Parker, 1962; Dungey, 1961; Gold and Hoyle, 1960) that instabilities may play an important role in explaining the origin of solar flares. Dungey (1961) has suggested that a pinch instability can take place in the ionized solar atmosphere near the neutral point of the sunspot magnetic fields and also that runaway electrons may develop in the transition region, leading to a two-stream instability. The possibility of the pinch instability is based on the assumption that a current may be driven along the neutral point (or line) between two regions of oppositely directed magnetic fields to such an intensity that the self-magnetic field of the current overpowers all other forces. Although this may be a possibility, our analysis in section 2 shows that such an instability does not arise in the equilibriums considered here. Gold and Hoyle (1960) consider a particular geometry of a complex sunspot group (in relation to the problem of origin of flares) in which a pair of bipolar sunspots is parallel to another pair of bipolar sunspots. We shall consider here the simpler geometry of four sunspots each of which belongs to a pair of conjugate spots arranged in a straight line. This geometry is considered following Sweet (1958).

Severny (1958, 1959) and Severny and Shabhanski (1961) have worked out a model of the pinch effect in the neighborhood of the neutral point of the sunspot fields assuming that strong currents are generated by the magnetic field gradients. Severny arrives at the conclusion that the plasma being compressed by the two sunspot fields is unstable when the magnetic pressure is greater than the kinetic pressure of the plasma. The result is that if a compression begins it will proceed at an ever-increasing rate; i.e., the pinch effect takes place. The phenomena that take place in the chromospheric or coronal plasma near the sunspot region are quite different from the pinch effect produced in the laboratory. Figure 1 shows an equilibrium configuration of the pinched plasma produced in the laboratory. Figure 2 shows one of the possible equilibrium configurations that can arise in the solar plasma before the occurrence of a flare; this case is considered in detail in section 2. The basic difference between the two cases is that, whereas in the first the magnetic field outside the plasma is a decreasing function of the

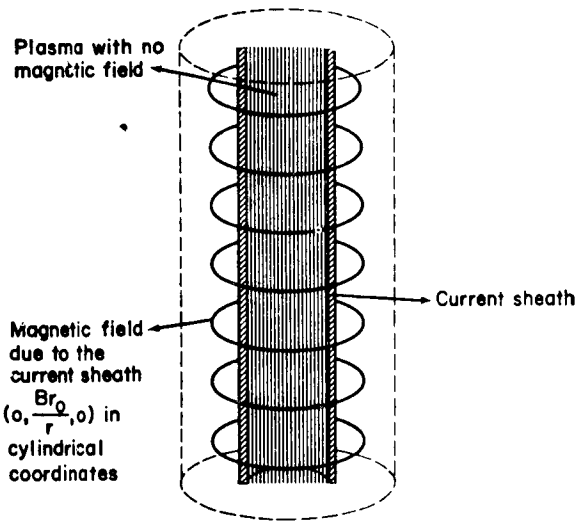


FIGURE 1.—Pinched plasma produced in the laboratory.

distance, in the second it is uniform. We shall show in section 2 that the equilibrium configuration considered in figure 2 is stable. The physical reason for the instability of the configuration in figure 1 is that any plasma moving away from the axis of the cylinder is at a point at which the magnetic field, and therefore the magnetic force, is less, and the perturbation continues to grow. Similarly, if any part of the plasma moves toward the axis of the cylinder it is further pushed by the magnetic field due to the increased magnetic force outside the plasma cylinder. If we now consider the equilibrium configuration of figure 2 we at once see that any displacement of the plasma in or out of its region is in the presence of the same magnetic force and there is no cause of instability as in laboratory pinches. The same holds for the equilibrium configuration sketched in figure 3.

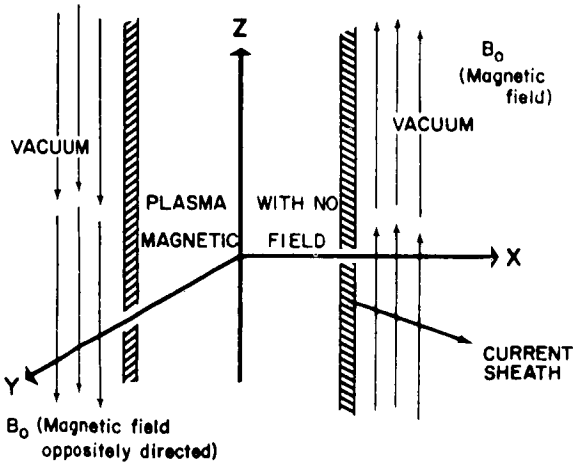


FIGURE 2.—A possible equilibrium configuration before the onset of flare. The magnetic field on both sides of a zero field region is oppositely directed. Outside is assumed vacuum. The plasma, the current sheath, and the magnetic field extend both ways, perpendicular to the plane of the paper, up to infinity. The current in the sheath flows in the Y direction.

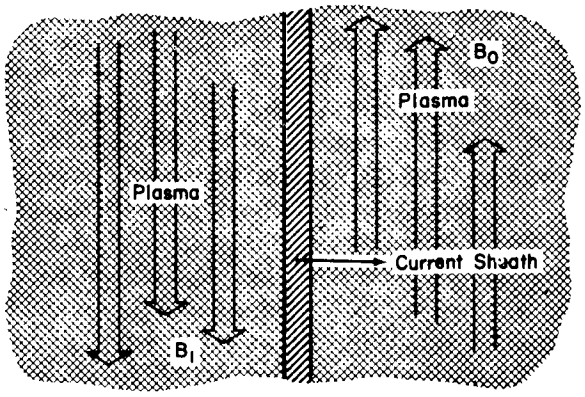


FIGURE 3.—Another possible hydromagnetic equilibrium of oppositely directed magnetic field regions. The plasma region and the current sheath extended to infinity in Y, Z directions. Right-hand half-plane is filled with plasma with magnetic field $(0, 0, B_0)$ in cartesian coordinates; the left half is filled with plasma with oppositely directed magnetic field $(0, 0, -B_1)$.

In sections 3 and 4 we shall analyze some of the known instabilities that apply in the present situation, with the understanding that an instability with growth rates of the right order constitutes the flare phenomena, as an instability helps to blow the material up. The discussion presented here does not help in understanding how the particles get accelerated to energies of the order of a few million electron volts; for this, acceleration processes must be considered separately.

2. MAGNETOHYDRODYNAMIC STABILITY OF EQUILIBRIUM CONFIGURATIONS

Observations show (Saverny, 1958) that, before the occurrence of a flare, the sunspots are approaching each other with the velocity of a few kilometers per second. Assuming that we have a group of four sunspots arranged in a straight line as considered by Sweet (1958, fig. 3), let us sup-

pose that a magnetohydrodynamic equilibrium is reached when the two magnetic fields of the sunspot groups are separated by a distance $2x_0$ with a region of zero magnetic field and hot plasma between. For simplicity, we neglect the gradient of the magnetic field on both sides of the zero field region. The possibility of such an equilibrium configuration will be increased if it is proved that the magnetic field gradients do not give rise to any serious instabilities. In section 3 we will show that the e folding growth time of any perturbation due to the magnetic field gradient instabilities is of the order $10^{10^{12}}$ seconds. This allows us to neglect any instabilities due to the gradients of the magnetic field. The pressure anisotropy instabilities do not arise in this case because the collision frequency is sufficiently high in the chromosphere as well as in the corona to preserve the pressure isotropy during the compression of the plasma. For coronal plasma of $N\sim10^8$ particles/cm^3, $T\sim10^5$ °K, the frequency of electron–electron collisions is of the order of 2.5×10^3 per second and the mean free path of an electron is of the order of 1 km. Since the time involved in the flare process is of the order of minutes, and the length is of the order of thousands of kilometers, the pressure will be assumed isotropic throughout the discussion.

The basic equations in this section are the magnetohydrodynamic equations as used by Kruskal and Schwarzschild (1954) and as further employed by Jaggi (1962). With the isotropic pressure the equations of motion are

$$\rho(dv/dt)+\text{grad}p=(1/c)\mathbf{j}\times\mathbf{B} \tag{1}$$

$$\partial\rho/\partial t=-\text{div}(\rho\mathbf{v}) \tag{2}$$

$$\partial\mathbf{B}/\partial t=\text{curl}(\mathbf{v}\times\mathbf{B}) \tag{3}$$

$$\frac{1}{p}\frac{dp}{dt}=\frac{\gamma}{\rho}\frac{d\rho}{dt} \tag{4}$$

$$\text{div}\mathbf{B}=0 \tag{5}$$

$$\text{curl}\mathbf{B}=(4\pi/c)\mathbf{j} \tag{6}$$

where the displacement current has been neglected. As a possible equilibrium configuration for the plasma let us assume that there is no magnetic field in the plasma confined between two parallel planes $x=\pm x_0$. For $x>x_0$ there is vacuum and the magnetic field (0, 0, B_0) is constant and parallel to the surface. Taking the z axis parallel to this magnetic field, we have a current in the y direction at $x=+x_0$. For $x<x_0$ the magnetic field is (0, 0, $-B_0$). The boundary conditions are

$$n\cdot[\mathbf{B}]=0 \tag{7}$$

$$[p+(B^2/8\pi)]=0 \tag{8}$$

If we perturb the boundary of the plasma-vacuum interface as

$$x=\pm x_0+i[(v_n)x=\pm x_0]/\omega \times\exp(ik_2y+ik_3z-i\omega t) \tag{9}$$

where the perturbations of the physical entities in the plasma are of the form (e.g., the perturbation in the pressure denoted by subscript 1)

$$p_1=p_1(x)\ \exp(ik_2y+ik_3z-i\omega t) \tag{10}$$

Linearizing equation 4 gives

$$p_1=\zeta^2\rho_1 \tag{11}$$

where $\zeta=(\gamma p_0/\rho_0)^{1/2}$ is the velocity of sound in the plasma, and p_0 and ρ_0 are the equilibrium pressure and density of the plasma, respectively. We easily find that

$$\mathbf{B}_1=0 \tag{12}$$

ρ_1 satisfies the equation

$$\frac{\partial^2\rho_1}{\partial x^2}-(k_1^2+k_2^2-\omega^2/\zeta^2)\rho_1=0 \tag{13}$$

and has for its solution

$$\rho_1=Ae^{Kx}+Be^{-Kx} \tag{14}$$

with $K=[k_1^2+k_2^2(\omega^2/\zeta^2)]^{1/2}$.

Also

$$p_1=\zeta^2(Ae^{Kx}+Be^{-Kx}) \tag{15}$$

$$\left.\begin{aligned} v_x&=-(i\zeta^2k/\omega\rho_0)(Ae^{Kx}-Be^{-Kx})\\ v_y&=(\zeta^2k_2/\omega\rho_0)(Ae^{Kx}+Be^{-Kx})\\ v_z&=(\zeta^2k_3/\omega\rho_0)(Ae^{Kx}+Be^{-Kx}) \end{aligned}\right\} \tag{16}$$

The solutions for perturbation in magnetic field

in the vacuum $x>x_0$ and $x<x_0$ are, respectively,

$$\left.\begin{aligned} B_{1x} &= DK_1 e^{-K_1 x} \\ B_{1y} &= Dik_2 e^{-K_1 x} \\ B_{1z} &= Dik_3 e^{-K_1 x} \end{aligned}\right\} \quad \text{and} \quad \left.\begin{aligned} B_{1x} &= D_1 K_1 e^{K_1 x} \\ B_{1y} &= D_1 ik_2 e^{K_1 x} \\ B_{1z} &= D_1 ik_3 e^{K_1 x} \end{aligned}\right\} \tag{17–18}$$

where $K_1=(k_2^2+k_3^2)^{1/2}$ and D and D_1 are constants of integration; the factor $\exp(ik_2y+ik_3z-i\omega t)$ has been omitted on both sides of equations 14 and 17–18 for convenience. Applying the boundary conditions (7) and (8) at $x=+x_0$ and $x=-x_0$, we obtain the dispersion equations

$$\left.\begin{aligned} &\omega^2(K_1/V_A^2k_3^2k)[\exp(2Kx_0)-1] \\ &\qquad =[\exp(2Kx_0)+1] \\ \text{or} \\ &\omega^2(K_1/V_A^2k_3^2k)[\exp(2Kx_0)+1] \\ &\qquad =[\exp(2Kx_0)-1] \end{aligned}\right\} \tag{19}$$

which show that ω^2 cannot be negative, and therefore the equilibrium is stable.

The other equilibrium configuration considered by us is shown in figure 3. For $x>0$, we have a uniform plasma of density ρ_0, pressure p_0, and the magnetic field $(0, 0, B_0)$. For $x<0$, we have a uniform plasma of density ρ_1, pressure p_1, and magnetic field $(0, 0, -B_1)$. At $x=0$, we have a current $(c/4\pi)(B_1-B_0)$ in the y direction. Following the analysis given above, we obtain the dispersion equation in the form

$$\frac{\left[k_2^2+k_3^2+\dfrac{[1-(\omega^2/k_3^2c_0^2)][1-(\omega^2/k_3^2V_0^2)]}{[1-(\omega^2/k_3^2c_0^2)-(\omega^2/k_3^2V_0^2)]}\right]^{1/2}}{\left[k_2^2+k_3^2+\dfrac{[1-(\omega^2/k_3^2c_0^2)][1-(\omega^2/k_3^2V_0^2)]}{[1-(\omega^2/c^2k_3^2)-(\omega^2/k_3^2V_1^2)]}\right]^{1/2}} = -\frac{[1-(\omega^2/k_3^2V_0^2)]}{[1-(\omega^2/k_3^2V_1^2)]} \tag{20}$$

where

$$c_0=(\gamma p_0/\rho_0)^{1/2} \qquad c_1=(\gamma p_1/\rho_1)^{1/2}$$
$$V_0=B_0/(4\pi\rho_0)^{1/2} \qquad V_1=B_1/(4\pi\rho_1)^{1/2}$$

If $\omega^2<0$, the left-hand side is positive and the right-hand side negative. This shows that no solution is possible for $\omega^2<0$; i.e., no instabilities exist. For $B_0=B_1$, $\rho_0=\rho_1$, $p_0=p_1$; i.e., the symmetrical case, the dispersion equation is of the form

$$k_2^2+k_3^2\frac{[1-(\omega^2/k_3^2c_0^2)][1-(\omega^2/k_3^2V_0^2)]}{[1-(\omega^2/k_3^2c_0^2)-(\omega^2/k_3^2V_0^2)]}=0 \tag{21}$$

which gives a quadratic equation having two positive roots. Thus there are no instabilities of the current layer in equilibrium between two ionized clouds having oppositely directed uniform magnetic fields. If we take the gradient of the magnetic field into account we predict that the current layer will still be stable because the magnetic field on both sides has increasing magnitude.

3. MAGNETIC FIELD GRADIENT INSTABILITIES

We shall now show that there are no serious magnetic field gradient instabilities for long wavelengths to affect the assumption of the equilibrium configurations studied in the previous section. Krall and Rosenbluth (1962) have considered a plasma in equilibrium in a magnetic field of the form

$$B=B_0(1+\epsilon x)i_z \tag{22}$$

where i_z is a unit vector in the z direction. This magnetic field configuration is quite suitable for our problem. The distribution function for the electrons and ions can be assumed to be

$$f_{0_j}=n(\alpha_j/\pi^2)(\beta_j/\pi)^{1/2}\exp[-\alpha_j(v_x^2+v_y^2)]\times\exp[-\beta_j(v_z^2)]g \tag{23}$$

where the subscript j refers to electrons ($j=e$) and ions ($j=i$); g satisfies the equation

$$(dg/dx)-(eB_0/mc)[1+\epsilon x](dg/dvy)=0 \tag{24}$$

Choosing

$$g_e=1-\epsilon^1[x+\tfrac{1}{2}\epsilon x^2-(vy/\Omega_e)]$$
$$g_i=1-\epsilon^1[n+\tfrac{1}{2}\epsilon x^2+(vy/\Omega_i)]$$

where $\epsilon^1=\epsilon B_0^2/[4\pi nk(T_i+T_e)]$, n being the number density of the particles, $\Omega_e=eB_0/mc$, $\Omega_i=eB_0/Mc$, $m=$ the mass of the electron and M that of an ion, Krall and Rosenbluth arrive at unstable solutions of the form e^{st} with the growth rate s given by

$$R_e(s)=\frac{2\left(\frac{T_e}{T_i}\right)^2\frac{\pi\Omega_i}{Xk^2R_1^2}E_i\left(X\frac{T_i}{T_e}\right)\exp\left(\frac{2XT_i}{T_e}\right)}{\{[E_i(XT_i/T_e)]^2+\pi^2\}^2} \tag{25}$$

for $kR_e<1$, $\epsilon^1/\epsilon\gg1$, and

$$R_e(s)=\frac{4\Omega_i\left(\frac{T_e}{T_i}\right)^{3/2}\exp\left(\frac{2XT_i}{T_e}\right){}_1F_1\left(\frac{1}{2};\frac{3}{2};\frac{XT_i}{T_e}\right)}{X(\pi X)^{1/2}\left\{\frac{4T_i}{\pi T_e}\left[{}_1F_1\left(\frac{1}{2};\frac{3}{2};X\frac{T_i}{T_e}\right)\right]^2+\frac{1}{X}\right\}^2} \tag{26}$$

for $kR_e>1$, $\epsilon_1/\epsilon\gg1$, where R_e and R_i denote the Larmor radii of electrons and ions respectively. Here $X=2d_i\Omega_i^2/k_e$ should not be confused with the coordinate x. k is the wave number of purely longitudinal electrostatic oscillations of the form

$$\mathbf{E}=(E\mathbf{k}/|k|)\epsilon^{i\mathbf{k}\cdot\mathbf{x}}E_i(XT_i/T_e)$$

$$\simeq\exp(XT_i/T_e)/(XT_i/T_e)$$

for large values of X, and ${}_1F_1$ is a confluent hypergeometric function. For the case of sunspot magnetic field gradients,

$$\epsilon=\frac{1}{B}\frac{dB}{dx}\simeq\frac{1}{10^{10}}\text{ cm}^{-1}$$

$$B_0\sim10^3\text{ gauss}$$

$$X\simeq0.243(B^2\lambda/T_e)$$

where $\lambda=2\pi/k$. For $T=5\times10^3$ °K, $\lambda\simeq1$ cm, $X\simeq10^{13}$, i.e., an extremely large quantity. Equation 25 yields

$$R_e(s)\simeq10_1{}^{21}/6\cdot10^{36}/e^{10^{12}}\sim10^{-10^{13}} \tag{27}$$

i.e., an extremely small number.

Conversely, if $kR_e>1$; i.e., wavelengths less than the electron cyclotron radius, the growth rate is extremely large, of the order of inverse of the growth rate for $kR_e<1$. This shows that only extremely small wavelengths are unstable. The Debye length in the lower chromosphere is $(kT/4\pi ne^2)^{1/2}\simeq5\times10^{-2}$ cm, and the electron cyclotron radius corresponding to 5×10^3 °K temperature is of the order of 10^{-3} cm. Any wavelengths less than the electron cyclotron radius will also be less than the Debye length, and the theory may not apply. (This is satisfied in the corona as well.) At any rate, the instability of extremely small wavelength, although taking place instantaneously, will not be efficient to enhance the thermal and electrical diffusion of the plasma.

The magnetic field gradient instabilities discussed so far in this section are only important much before the equilibrium between the magnetic pressure and the kinetic pressure is reached (if it is reached at all) because of the condition $\epsilon^1/\epsilon\gg1$ imposed on solutions 25 and 26. After the compression continues, the temperature of the plasma builds up; the pressure P begins to increase until it is comparable with the magnetic pressure $B^2/8\pi$. The kinetic pressure develops more rapidly than the magnetic field pressure (Severny, 1958, p. 321), and an equilibrium configuration can be reached when P is comparable with $B^2/8\pi$.

It might be thought that the pressure anisotropy instabilities can develop in this case, but it can be shown that owing to the slow rate of approach of the two sunspots and owing to the high collision frequency the pressure anisotropy cannot develop to the extent of producing an instability. As has been mentioned earlier, the collision frequency in the lower corona is of the order of about 10^6 per second, and, since the compression time is of the order of hundreds of seconds, the collisions will keep the pressure quite isotropic in the present situation. It is only if the phenomena take place in about 10^{-4} second that the pressure anisotropy can be important. To calculate the magnetic field, the density, and the temperature after the compression we may use the adiabatic equations deduced by Chew, Goldberger, and Low (1956). With $P_{||}=P_{\perp}$, these equations are

$$PB^2/N^3=\text{constant} \tag{28}$$

$$P/NB=\text{constant} \tag{29}$$

from which we obtain

$$B^3/N^2 = \text{constant} \tag{30}$$

If L_0, B_0, P_0 denote the initial length, magnetic field density, and the pressure of the plasma, the quantities $B_1N_1P_1$ for a given L_1 can be calculated from $N_1L_1 = N_0L_0$, $B_1^3/N_1^2 = B_0^3/N_0^2$, $P_1/N_1^{5/3} = P_0/N_0^{5/3}$. For $B_0 = 10^3$ gauss, $N_0 = 10^{12}/\text{cm}^3$, $L_0 = 10^{10}$ cm, $T_0 = 5\times10^3$ °K, we obtain for $L_1 = 10^7$ cm, $N_1 = 10^{15}/\text{cm}^3$, $B_1 = 10^5$ gauss, $T_1 = 5\times10^5$ °K. The collision frequency remains about the same as before the compression.

4. RESISTIVE INSTABILITIES

The instabilities discussed so far apply in the case of infinite conductivity of the plasma under consideration. Recently some work has been done assuming finite conductivity of the plasma (Furth et al., 1963). We shall use the results to make an estimate of the growth rates of resistive instabilities.

The infinitely thin current layer as sketched in figure 3 will now be assumed to have finite thickness denoted by a. We will also assume a magnetic field in the current layer. With finite conductivity, equation 3 has to be replaced by

$$\partial\mathbf{B}/\partial t = \text{curl}(\mathbf{v}\times\mathbf{B}) - \text{curl}[(\eta/4\pi)\,\text{curl}\mathbf{B}] \tag{31}$$

where η is the isotropic resistivity of the plasma. The magnetic field in the current sheath has the form

$$\mathbf{B} = i_z B_z(x) \tag{32}$$

Further assuming the current layer to be incompressible (this amounts to using $\text{div}\mathbf{v} = 0$ in place of equation 2 and deleting equation 4), two unstable modes are deduced by Furth, Killeen, and Rosenbluth for waves propagating in the z direction. They are called the rippling mode and the tearing mode. The rippling mode arises in the case in which the resistivity varies with distance in the current layer; the growth rate of the instability which is the inverse of the e folding time of a perturbation is given by

$$R_e(s) = \alpha^{2/5}S^{2/5}/4\pi a^2\sigma \tag{33}$$

where $\sigma(=1/\eta)$ is the conductivity of the plasma in emu

$$S = 10^{-2}[T^2/(4\pi P_0/B_0)^{1/2}]$$

T in electron volts, $\alpha = ka$, where k is the wave number of waves in the current layer (propagating in the current layer). Also this instability occurs for $S^{-2/7} < \alpha < S^{2/3}$. Using $\sigma = 2\times10^{-14}\ T^{3/2}$ emu, $T = 10^6$ °K, $a = 10^5$ cm (10^6 °K$\cong$86 ev), we have $S = 7\times10^3$, and the e folding time which is the inverse of $R_e(s)$ is of the order of $10^3/k^{2/3}$. Now, from $S^{-2/7} < \alpha < S^{2/3}$ we get $2.8 < 1/k < 10^6$. For long wavelengths, the e folding time for the rippling mode is large. Only for small wavelengths, of the order of a few centimeters, the times are reasonable and the instability of rippling mode is important for the problem in hand.

The tearing mode is unstable for $\alpha < 1$ or wavelengths greater than the thickness of the current sheath. In this case the e folding time is of the order $10^6\ k^{2/5}$, where $k \gtrsim 1/2500$. This time will be small for long wavelengths, and the order of magnitude is also suitable for the problem.

The above results are derived on the assumption that the current layer is incompressible, and it is proved by Furth, Killeen, and Rosenbluth

TABLE 1.—*Rippling and Tearing Mode Instability*

[If t_r denotes the e folding time for the rippling mode, this table evaluates $t_r(k)^{2/5} = 2.96\times10^{-2}T^{9/10}/B_0^{2/5}$. The e folding time for the tearing mode is obtained from $t_t = t_r10^4(k)^{4/5}$]

T, °K	$B=10$ gauss	$B=10^2$ gauss	$B=10^3$ gauss	$B=10^4$ gauss	$B=10^5$ gauss	$B=10^6$ gauss
10^4	44.4	18.6	7.42	2.96	1.17	0.469
10^5	3.7×10^2	1.48×10^2	59	23.5	9.35	3.72
10^6	2.96×10^3	1.17×10^3	4.69×10^2	1.78×10^2	74.03	29.6
10^7	3.72×10^3	1.47×10^3	5.9×10^2	2.24×10^2	93.02	36.58

that the results are not much different if compressibility is taken into account.

To arrive at the above figures, we have used the temperature of 10^6 °K in the higher chromospheric region. Some people (De Jager, 1959, 1961) are of the opinion that the temperature of the chromosphere region does not rise very much and that only the density becomes higher. A differing view, based on theoretical models, is adopted in section 3 of this article, following Severny (1958). For this reason we have evaluated the growth rates of rippling and tearing modes in table 1. The magnetic field at the neutral point, where the current sheath is formed, will be weak. This is another reason why the growth rate can drop off by several orders of magnitude.

For the rippling mode the e folding time is obtained by multiplying each member of the table by $(k^{-1})^{2/5}$. We find that

$$\begin{matrix} 6.7 \\ 30.7 \end{matrix} > (k^{-1})^{2/5} > \begin{matrix} 1 \\ 10^{-3} \end{matrix}$$

for the whole range of the table. The upper numbers apply for $B=10$, $T=10^4$; the lower, for $B=10^6$, $T=10^7$. An intermediate value of 1 shows that suitable time scales are obtained for almost all the values of B and T in the table. The above results are deduced for $N\sim10^{15}$ particles/cm^3. If, instead, we use $N\sim10^{12}$ particles/cm^3, it amounts to multiplying each number in the table by $10^{3/5}\sim4$. N occurs in t_r or t_t as $N^{-1/5}$.

For the tearing mode the e folding time is obtained by multiplying each member of the table by $10^4(k)^{2/5}$, where, for the whole range of the table,

$$10^{-2} > (k)^{2/5} > \begin{matrix} 0.0208 \\ 0.0023 \end{matrix}.$$

The upper number applies for $B=10$, $T=10^4$; the lower, for $B=10^6$, $T=10^7$. Assuming $k^{2/5}$ to be of the order of 10^{-3}, we see that the e folding times, for large magnetic fields and small temperatures, are reasonable. We therefore conclude that the resistive instabilities give reasonable times for the problem in hand.

ACKNOWLEDGMENTS

I should like to thank Drs. J. M. Burgers, W. N. Hess, E. N. Parker, and D. A. Tidman for stimulating discussion.

I was a NAS Research Associate during this work.

REFERENCES

Chew, G. F., Goldberger, M. L., and Low, F. E., An attempted derivation of a one-fluid magnetohydrodynamics, *Proc. Roy. Soc. London, A*, **236,** 112 (1956).

De Jager, C., Structure and dynamics of the solar atmosphere, *Handbuch der Phys.*, **52,** 80, Springer, Göttingen (1959).

———, The development of a solar center of activity, *Vistas in Astronomy*, edited by A. Beer, vol. 4, pp. 143–183, Pergamon Press (1961).

Dungey, J. W., *Kyoto Conf. Cosmic Rays and Earth Storms (IAU)*, Kyoto, Japan (1961).

Furth, H., Killeen, J., and Rosenbluth, M. N., Finite resistive instabilities of a sheet pinch, to be published in *Phys. Fluids* (1963).

Jaggi, R. K., Magnetohydrodynamic stability of a self-gravitating plasma, *Z. Astrophys.*, **54,** 190 (1962).

Gold, T., and Hoyle, F., On the origin of solar flare, *Monthly Notices Roy. Astron. Soc.*, **122,** 89–105 (1960).

Krall, N., and Rosenbluth, M. N., Trapping instabilities in a slightly inhomogeneous plasma, *Phys. Fluids*, **5,** 1435 (1962).

Kruskal, M., and Schwarzschild, M., Some instabilities of a completely ionized plasma, *Proc. Roy. Soc. London, A*, **223,** 348 (1954).

Parker, E. N., Theory of reconnection and annihilation of magnetic fields with application to solar flares and interstellar space, unpublished (1962).

Severny, A. B., Nonstationary processes in solar flares as a manifestation of the pinch effect, *Soviet Astron. (A.J.)*, **2,** 310 (1958).

———, On the appearance of cosmic rays in the pinch effect in solar flares, *Soviet Astron. (A.J.)*, **3,** 223 (1959).

———, and Shabhanski, V. P., The generation of cosmic rays in flares, *Soviet Astron. (A.J.)*, **4,** 583 (1961).

Sweet, P. A., The neutral point theory of solar flares, *IAU Symp. on Electromagnetic Phenomena in Cosm cal Phys.*, **6,** 123 (1958).

A THEORETICAL MODEL FOR SUNSPOT COOLNESS

R. K. JAGGI

Goddard Space Flight Center

We have developed a theoretical model for the behavior of sunspots as individuals. Two models, called the current sheath model and the snowplow model, known in the theory of the pinch effect are worked out for the present problem of the sunspot expansion. Using the observational fact that the magnetic field of the sunspots grows to about 3,000 gauss in about 10 days, numerical calculations based upon the current sheath model show that the sunspot area grows with the magnetic field and begins to pulsate when the magnetic field stops growing. The amplitude and frequency of oscillation depend upon the mass in the current sheath and the maximum magnetic field. The model suggests that the bright ring about the sunspots may exist at chromospheric height where the density of the material is about 10^9 particles/cm^3.

INTRODUCTION

Among the unsolved problems in solar physics reviewed by Goldberg and Dyer[1] is the difficulty in obtaining a theoretical interpretation of the mechanism of sunspot cooling. They proposed two explanations: The gas pressure in the sunspot is low because part of the total pressure is exerted by magnetic forces; in that case the configuration must approach an equilibrium in which the sum of the kinetic and magnetic pressures equals the kinetic pressure outside. The cooling could also be due to the forced expansion of a rising gas column. This second possibility is discarded because there is no observational evidence that the gas column is really rising. Many other models used to explain the coolness are based on the presence of a strong magnetic field (reference 2, p. 172). Biermann[3] assumed that the low temperature is probably maintained by the strong magnetic fields inhibiting convective transport of energy, thus causing steep temperature gradients in the outer parts of the spots.

At this point it would be instructive to estimate the time in which the sunspot region, radiating like a blackbody, would fall from 6500° to 4500°K, the temperatures of the photosphere and a typical sunspot respectively. If convection is completely stopped, the time of this temperature fall is determined by

$$L \frac{\partial}{\partial t} (nkT) = \sigma T^4 ,$$

where $\sigma = 5.6 \times 10^{-5}$ erg/cm^2-sec-deg^4, and L, the total depth of the cool region, is given by

$$L \frac{nk(6500 - 4500)}{t} \approx \sigma (5000)^4 ,$$

with k the Boltzmann constant and σ the radiation constant. For $L = 10^8$ cm and $n = 10^{17}$, the above equation gives $t \approx 7.8$ sec, which shows that the region will cool down instantaneously. Some work has been done on the inhibition of convection in the presence of a magnetic field.[4,5] The difficulty with Biermann's explanation is twofold: (1) it fails to explain the growth of the sunspot area with time; and (2) the cooling appears 1 or 2 days *after* the observation of the magnetic field. Also, any explanation of the sunspot coolness must at the same time explain the Evershed effect as well as the bright ring around the sunspots (as known by observation).

De Jager[2] has mentioned that the ions with energies greater than 3 ev are reflected by the sunspot's magnetic field; but this is only effective in the upper, less dense part of the solar atmosphere and therefore does not explain lower temperature at the photospheric level. The particles

reflected by the strong magnetic field of the sunspot will leave the region only if the mean free path is large enough to permit their escape. From table 1 (in the last section of this paper), it is clear that the mean free path is large only in the upper chromosphere.

OBSERVATIONAL DATA

Before considering the theoretical model, we shall present some observational data. Most of this material was given in two articles by De Jager.[2,6]

The development of sunspots forms part of the development of a *center of activity* (CA). De Jager[6] has divided the development of a CA into four parts: (1) prespot phase; (2) spot phase; (3) bipolar magnetic postspot phase; and (4) unipolar magnetic phase. Of these, the second is of most interest in the present problem. A typical sunspot consists of an umbra and a penumbra, where the latter shows a radial filamentary structure, being slightly brighter than the former. The mean umbral area of large sunspots is of the order of 5×10^{-4} of the sun's visible hemisphere and the radius of the umbra is of the order of 2.2×10^{9} cm.

In the prespot phase the development of a bipolar magnetic field region as well as faculae are observable without the spots. The limits of the faculae coincide more or less with the sunspot region. The spot appears 1 or 2 days after the appearance of the magnetic field. The spot and

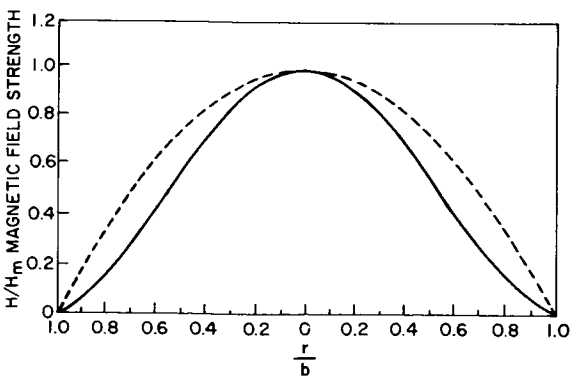


FIGURE 1.—Variations of magnetic field strength over a sunspot as function of radius from the center of the spot. The distance to the spot center is expressed in terms of penumbral radius.

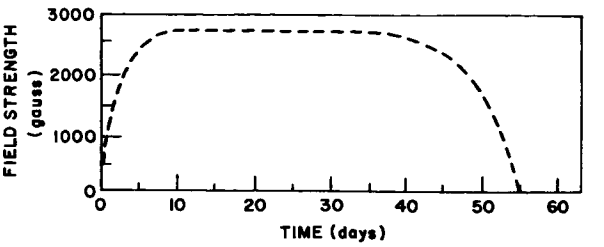


FIGURE 2.—Variation of the field strength versus time in days for a 55-day spot.

the magnetic field region increase in area simultaneously.

15-day time sequence of the development of a typical sunspot group is given by De Jager (reference 2, p. 167). On the first day a single spot or a spot group without penumbra or without bipolar structure is observed on the sun. On the second day the bipolar group and penumbras of some of the bipolar spots appear. From 5 to 10 days great bipolar spots as well as many small ones appear. From 10 to 15 days great bipolar spot groups remain without their small companions. During this time unipolar spots may also be observable.

Kiepenheuer[7] quotes Thiessen[8] as giving the following radial distribution of the vertical magnetic field in a sunspot:

$$B_z(r)\approx B_m[1-(r^2/b^2)], \qquad (1)$$

where b is the radius of the outer edge of the penumbra, the subscript z denotes the z component of the field and the subscript m the field at the center of the spot, and r the radial distance measured from the axis of the spot. However, the magnetic field distribution as determined by Mattig[9] is of the form:

$$B_z(r)=B_m[1-(r^4/b^4)]\exp(-2r^2/b^2).$$

Figure 1 is an approximate representation of this field (taken from reference 2, p. 156). At the center of the spot the magnetic field B_m is of the order of a few thousand gauss. Assuming B_m to be 10^3 gauss we find that the magnetic pressure $B_2/8\pi$ is approximately 4×10^5 dynes/cm^2 and is much higher than the kinetic pressure of the solar photospheric plasma $nkT\approx3\times10^3$. Thus the sunspot is under the action of strong magnetic forces. Equating $B^2/8\pi$ with the kinetic pressure 3×10^3, we obtain $B\approx260$ gauss. Therefore whenever the

magnetic field is comparable to or higher than this value the sunspot will be under the action of strong magnetic forces.

The time variation of the magnetic field observed over the sunspots is presented in figure 2, which is taken from Cowling.[10] Analytically the curve may be approximated by

$$B_m = 3000[1 - \exp(-t/2)]\{1 - \exp[(t/5.5) - 10]\} \quad (2)$$

where t is measured in days.

We shall now discuss the adiabatic expansion of sunspots for the two models known as the current sheath and the snow plow models, in order to determine the rate of cooling in both instances.

THE CURRENT SHEATH MODEL

The basic equations of magnetohydrodynamics are

$$Nm_i \frac{dv}{dt} = \frac{1}{c} \mathbf{j} \times \mathbf{B} - \text{grad } p \,, \quad (3)$$

$$\frac{\partial N}{\partial t} = -\text{div}(Nv) \,,$$

$$\text{Curl } \mathbf{B} = \frac{4\pi}{c} \mathbf{j} \,, \quad (4)$$

$$\text{div } \mathbf{B} = 0 \,,$$

$$\frac{\partial \mathbf{B}}{\partial t} = \text{Curl } (\mathbf{v} \times \mathbf{B}) \,,$$

$$\text{div } \mathbf{E} = 4\pi\epsilon \,,$$

where the symbols have their usual meaning, ϵ is the charge density, and m_i the mass of an ion. From equations 1 and 4, we obtain, in cylindrical coordinates

$$\mathbf{j} = (cB_m/4\pi)[0, 2r/b^2, 0];$$

therefore the force,

$$c^{-1}\mathbf{j} \times \mathbf{B} = \{B_m r/2\pi b^2[1 - (r^2/b^2)], 0, 0\},$$

is directed radially outward force. Consequently the cylinder must expand in the radial direction.

To simplify the model we modify the distribution of the magnetic field given by equation 1 in figure 1 to that given in figure 3. A portion of the cylinder is assumed to have a uniform magnetic field while the gradients of the magnetic field are assumed to be in the rest of the cylinder.

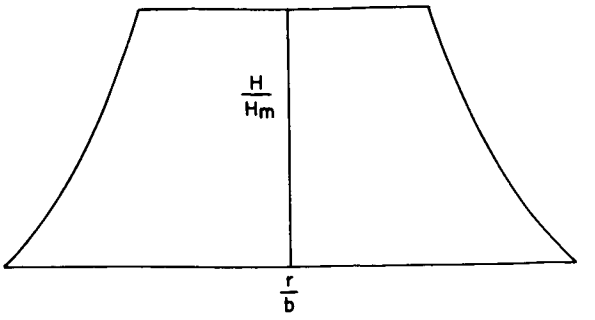


FIGURE 3.—Modified form of figure 1 used in calculating the expansion of sunspots.

The region occupied by this current is called "the current sheath." Equation 3 can be rewritten as

$$Nm_i \frac{dv}{dt} = -\frac{1}{8\pi} \text{grad } B^2 + \frac{1}{4\pi} (\mathbf{B} \cdot \nabla) \mathbf{B} - \text{grad } p \,.$$

In the present case the term $(\mathbf{B} \cdot \nabla)\mathbf{B} = 0$ because it is assumed that B is in the z-direction but there are no gradients in the z-direction. Integrating over the thickness of the current sheath, multiplying both sides by $2\pi r$, assuming that the acceleration over the cross-section of the current sheath and the density in the current sheath are uniform, we obtain

$$\frac{M}{\alpha} \frac{d^2 r}{dt^2} = 2\pi r \left(\frac{\bar{B}_i^2}{8\pi} + \bar{p}_i - p_0 \right), \quad (5)$$

where p_0 is the kinetic pressure due to the gases outside the cylinder, $\bar{p}_i$ the average kinetic pressure due to gases within the cylinder, and $\bar{B}_i$ the average magnetic field within the cylinder, and M the total mass contained in the cylinder. α has been introduced to insure that only a fraction of the total mass of the cylinder is contained in the current sheath. We shall assume that p_0 remains constant and that $\bar{p}_i = p_0$ at $t = 0$. The changes in

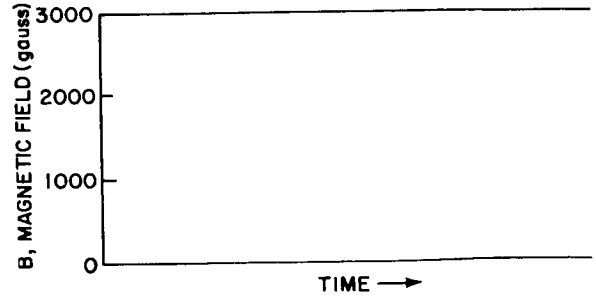


FIGURE 4.—Profile used to deduce equation 10.

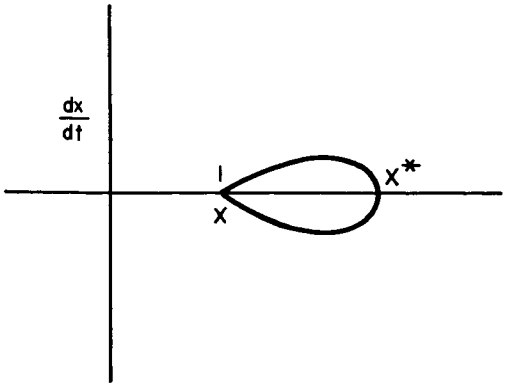


FIGURE 5.—Representation of equation 19.

the kinetic pressure satisfy the ordinary adiabatic law

$$\frac{p}{N^{5/3}} = \text{Constant} . \qquad (6)$$

If B_0 is the initial magnetic field and r_0 the initial radius of the cylinder, the constancy of the magnetic flux in the cylinder gives

$$\pi r_0^2 B_0 = \pi r^2 \bar{B}_i ,$$

or

$$\bar{B}_i = B_0 \left(\frac{r_0}{r}\right)^2 . \qquad (7)$$

Then, since N is proportional to $1/r^2$, equation 6 yields

$$\bar{p}_i = p_0 \left(\frac{r_0}{r}\right)^{10/3} . \qquad (8)$$

Substituting equations 7 and 8 into equation 5 gives

$$\frac{M}{2\pi a} \frac{d^2 r}{dt^2} = r\left[\frac{B_0^2}{8\pi}\left(\frac{r_0}{r}\right)^4 + p_0\left(\frac{r_0}{r}\right)^{\frac{10}{3}} - p_0\right] .$$

Now by writing $r/r_0 = x$, $B_0^2/8\pi p_0 = \beta$ and $(m/2\pi\alpha p_0)^{1/2} t' = t$ the above equation becomes

$$\frac{1}{x} \frac{d^2 x}{dt'^2} = \frac{\beta}{x^4} + \frac{1}{x^{\frac{10}{3}}} - 1 . \qquad (9)$$

To discuss the solution of equation 9, the dependence of B on time must be known. According to observations, equation 2, and figure 2, the magnetic field rises in a matter of 10 days to a

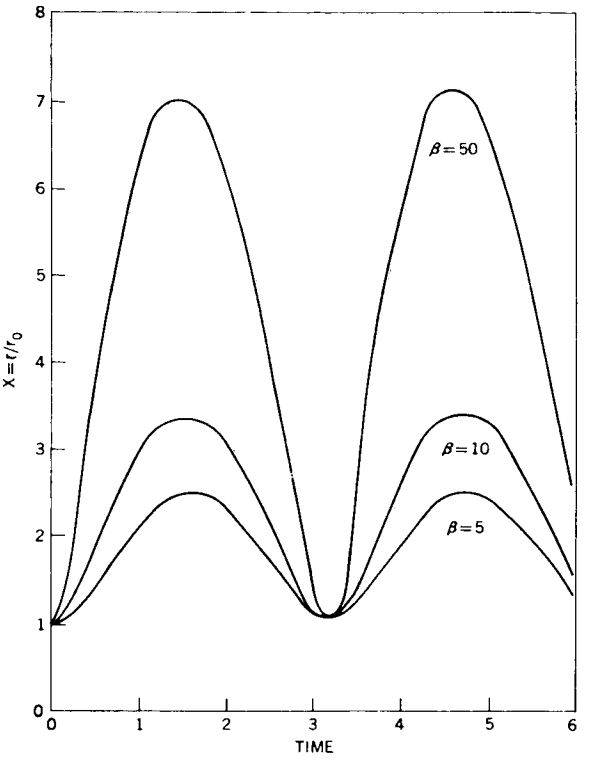


FIGURE 6.—The solution of equation 10 for $\beta = 5$, 10, 50.

value of about 3000 gauss and remains at that value for about 40 days. Since the growth time is comparatively shorter than the lifetime of the magnetic field we may assume the magnetic field as a function of time given by figure 4.

By assuming this magnetic field profile, it is possible to integrate equation 9. The integration yields

$$\frac{1}{2}\dot{x}^2 = C - \frac{1}{2}x^2 - \frac{3}{4}x^{-\frac{4}{3}} - \frac{1}{2}\beta x^{-2}$$

If $t' = 0$, $x = 1$, $\dot{x} = 0$ can be assumed as the initial condition, then

$$\dot{x}^2 = (1 - x^2) + \frac{3}{2}\left(1 - x^{-\frac{4}{3}}\right) + \beta\left(1 - x^{-2}\right) , \qquad (10)$$

where the dot denotes differentiation with respect to t'. Figure 5 shows the relationship of $\dot{x}$ and x; and $x = 1$, $x = x^*$ are the solutions of

$$1 - x^2 + \frac{3}{2}\left(1 - x^{-\frac{4}{3}}\right) + \beta\left(1 - x^{-2}\right) = 0 .$$

Initially x is an increasing function of t because

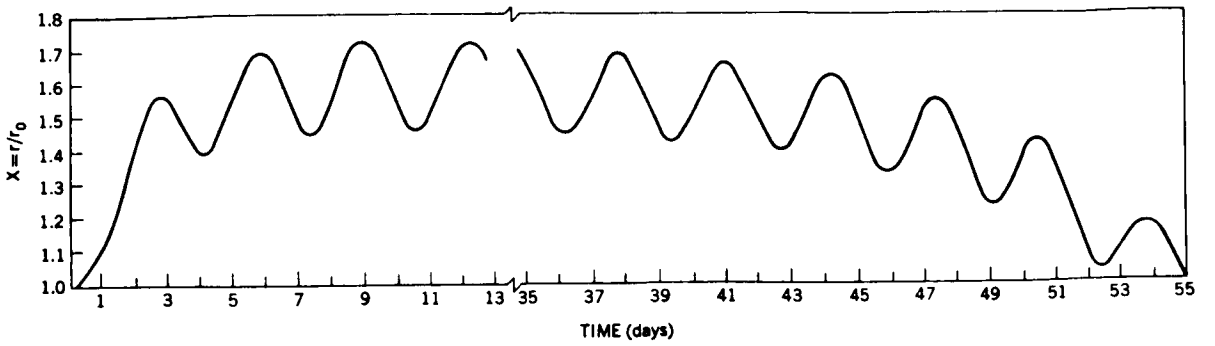


FIGURE 7.—A 55-day spot represented by equation 11 for $\zeta=5$.

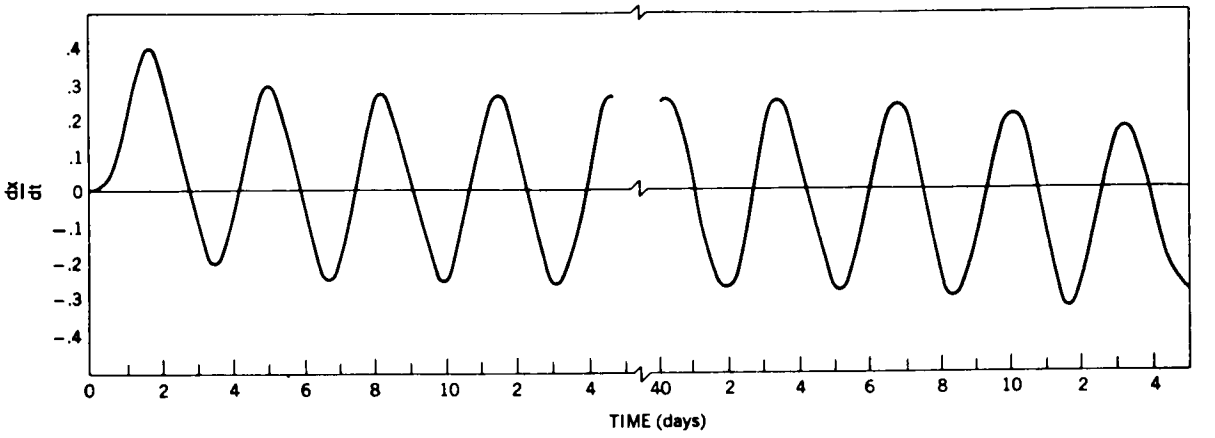


FIGURE 8.—Representation of equation 11 for $\zeta=5$.

$\ddot{x}>0$ at $t=0$. Figure 5 shows that x is an oscillatory function of t. A complete solution of equation 9, obtained on IBM 7090 computer, is represented in figure 6 for $\beta=5$, 10, 50. Here, the amplitude of oscillation increases with β.

To obtain a solution of equation 9, in the general case, where the profile of the magnetic field is given by equation 3 we must be cautious. If we use the expression for $\bar{B}_i$ given in equation 3 the solution is unstable because $\ddot{x}>0$ for all t and for any reasonable value of the magnetic field strength. For this reason we assume that the magnetic field is induced into the area at the rate given by equation 3. The problem here is to find a magnetic field profile which gives $B_0(t)/x^2$ close to the curve given in figure 2 and which gives a reasonably good fit for the area variation. Assuming then that

$$B_0(t) = B\left(1 - e^{-\frac{t}{2d}}\right)\left(1 - e^{\frac{t}{5.5d} - 10}\right),$$

where t is now measured in seconds and d is the number of seconds in one day, we have—instead of equation 9

$$\frac{d^2x}{dt^2} = -x + x^{-\frac{7}{3}} + \frac{\zeta}{x^3}\left(1 - e^{-\frac{t}{2t_1}}\right)^2\left(1 - e^{+\frac{t}{5.5t_1} - 10}\right)^2, \qquad (11)$$

where

$$t_1 = \frac{d}{\sqrt{\frac{M}{2\pi a p_0}}}, \quad \zeta = \frac{B^2}{8\pi p_0}.$$

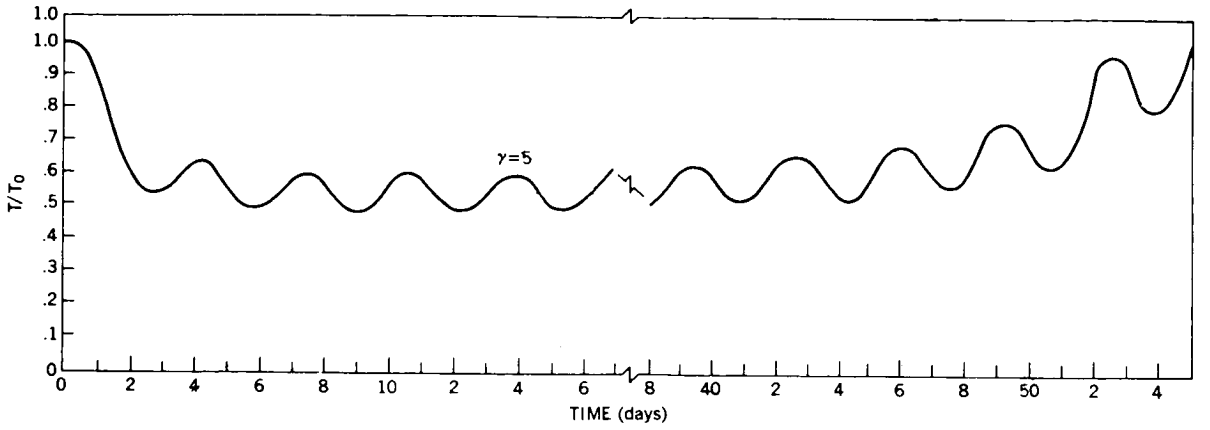


FIGURE 9.—T/T_0 versus t represented by equation 11 for $\zeta=5$.

A solution of equation 10 for $\zeta=5$, $t_1=1$ is presented in figure 7. The radius of the cylinder reaches a value of 1.6, oscillates about it and finally returns to its original value when the magnetic field has disappeared. The period of oscillation is 3 days and the amplitude of oscillation is about $0.25r_0$. Figures 8 and 9 present the velocity dx/dt and the temperature ratio T/T_0 as a function of t.

The assumption $t_1=1$ corresponds to $M/2\pi\alpha p_0=(86{,}400)^2$ or $\alpha\approx 6\times 10^3$ where we have used $r_0=10^9$ cm and $p=10^4$. Therefore, a fraction $1/(6\times 10^3)$ of the total mass of the cylinder is contained in the current sheath. For much higher values of α, the value of t_1 becomes large and the period of oscillation of the cylindrical sheath may become much smaller than that of figure 7.

THE SNOW PLOW MODEL

In the snow plow model, first discussed by Rosenbluth,[11] the magnetic field is homogeneous throughout the interior of the cylinder. The current sheath is in an infinitely thin layer of negligible mass. As the sheath expands it collects all the mass with which it comes in contact, as a snow plow collects the snow in its path. If r_0 is the initial radius of the cylinder and r the radius at any time t, the mass with the sheath is $\pi N m_i(r^2-r_0^2)$ where N is the average number density of ions of mass m_i. Newton's equation of motion of the sheath then becomes

$$\frac{d}{dt}\left[\pi N m_i\left(r^2-r_0^2\right)\frac{dr}{dt}\right]=2\pi r\left(\frac{B_i^2}{8\pi}+\bar{p}_i-p_0\right).$$

Using the transformations

$$\frac{r}{r_0}=x,$$

$$\beta=\frac{B_0^2}{8\pi p_0},$$

$$\left[\frac{N m_i r_0^2}{2p_0}\right]^{\frac{1}{2}} t'=t,$$

we obtain the equation

$$\frac{d}{dt'}\left[\left(x^2-1\right)\frac{dx}{dt'}\right]=x\left(\frac{\beta}{x^4}+x^{-\frac{10}{3}}-1\right), \quad (12)$$

where equations 7 and 8 were used to express B_i in terms of B_0 and $\bar{p}_i$ in terms of p_0. With the magnetic field profile of figure 4, this equation was solved on an IBM 7090 and its solution is shown graphically in figure 10. The initial conditions satisfied by this equation are $t'=0$, $x=1$, $\dot{x}=$

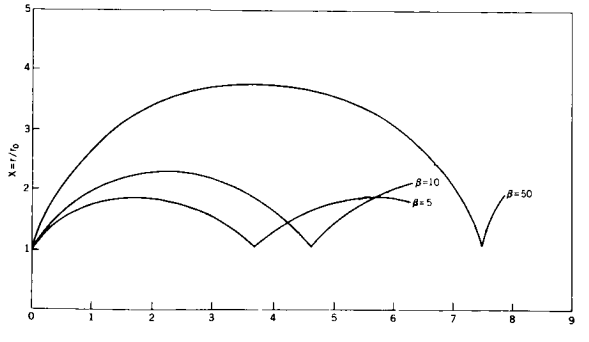


FIGURE 10.—The solution of equation 12 for $\beta=5$, 10, 50.

$(\beta/2)^{1/2}$. The solution is shown in figure 10 for $\beta=5$, 10, and 50.

A peculiar feature of this differential equation is that while in the expanding phase the sheath collects all mass with which it comes in contact; however, in the contracting phase it is losing its mass—not exactly in the same way as a snow plow because the snow plow loses its mass instantaneously when the direction of the velocity is reversed. We can therefore only follow the calculation of the snow plow model until the velocity reverses in sign. For that reason this model will not be discussed in detail.

THE EFFECT OF DENSITY AND PRESSURE GRADIENTS ON THE DYNAMICS OF SUNSPOTS

The pressure and density both rise in the lower levels of the photosphere. Figures 11 and 12 show the graph of the variation of the density P and the pressure p from the surface of the photosphere to the depth of a few hundred kilometers. In the solar photosphere, the pressure rises more rapidly with depth than the density. For example

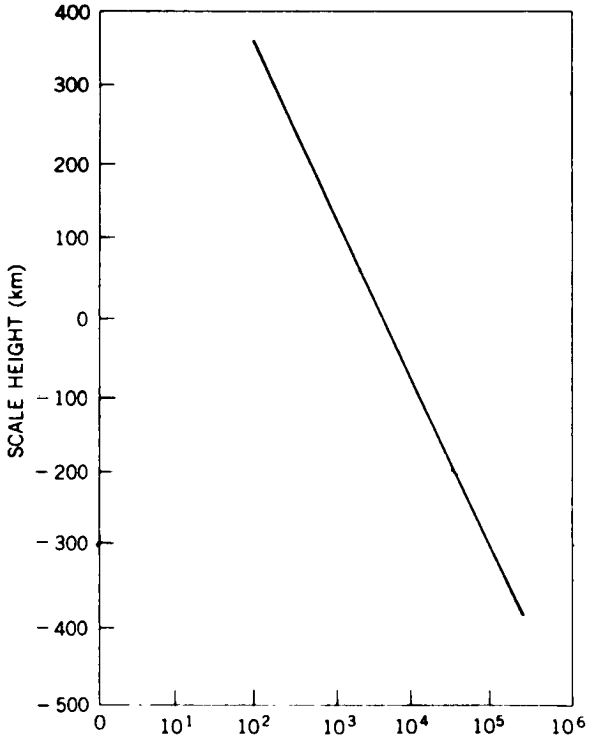


FIGURE 11.—Variation of ρ versus height in the solar photosphere (taken from reference 10, p. 127, table II).

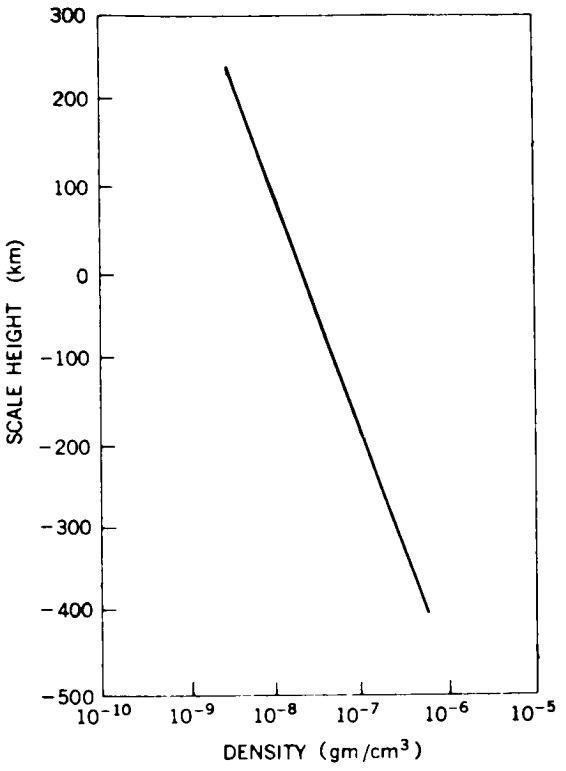


FIGURE 12.—Variation of p versus height in the solar photosphere (taken from reference 10, p. 127, table II).

at a depth of about 400 km the pressure is about 20 times that at the surface of the photosphere and the density is about 10 times that at the surface. If the photosphere is assumed to be stratified, the lower layers have less amplitude of oscillation than those at the surface. Also if the lower layers oscillate independently the frequency of oscillation can be different at the lower levels. Replacing p_0 with $20p_0$ in equations 5 and 8, we obtain in place of equation 9:

$$\frac{1}{x}\frac{d^2x}{dt'^2} = \frac{\beta}{x^4} + \frac{20}{x^{\frac{10}{3}}} - 20 . \qquad (13)$$

The solution of this equation is shown in figure 13.

The frequency of oscillation of x in figure 13 is about 0.75, and in figure 6 it is about 3.2. Thus the two periods of oscillation are

$$t_1 = 3.2\left(\frac{\pi r_0^2 \rho'}{2\pi a p_0}\right),$$

$$t_2 = 0.3\left(\frac{\pi r_0^2 \rho}{2\pi a p_0}\right).$$

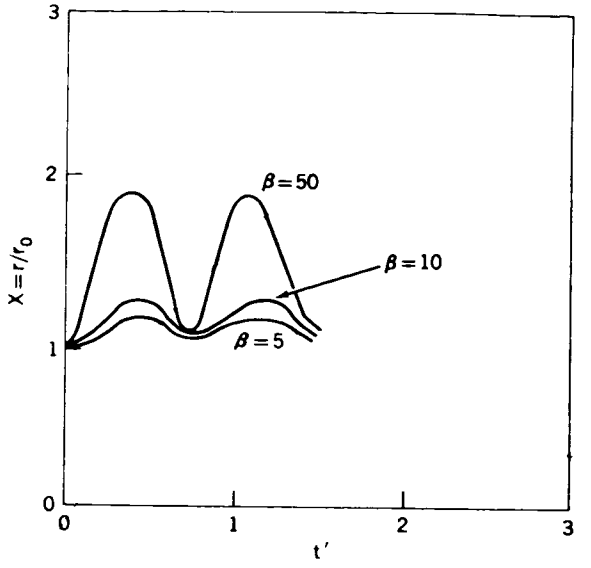


FIGURE 13.—Representation of equation 13 for $\beta=5$, 10, 50.

Therefore the ratio $0.75/3.2=(\rho'/\rho)^{1/2}$ is greater than or less than one depending on whether $\rho' \lesseqgtr 18.2\rho$.

In order to draw any conclusions about the frequency of oscillation we must know the variation of density with depth. From the figures given by Minneart (1953), in figure 12, we find that $\rho' \simeq 10\rho$. We therefore conclude that the period of oscillation at lower levels in the solar photosphere is approximately the same as at the surface. This subject will be taken up in a later paper.

THE BRIGHT RING AROUND SUNSPOTS

During the initial growth of the 55-day spot considered the velocity of the current sheath has a value of about 1–10 km/sec. Charged particles with their thermal velocities suffer reflection from these initially outward moving sheath and gain energy. These particles will then move a few mean free paths before loosing the energy gained from reflection. Therefore additional light in the solar photosphere or chromosphere, will be generated and a bright ring around a spot, with a width of the order of a mean free path will be created.

The self-collision time of particles of mass m and charge e is given by

$$t_c = \frac{m^{\frac{1}{2}}(3kT)^{\frac{3}{2}}}{8 \times 0.714\pi me^4z^4 \log \Lambda},$$

where

$$\Lambda = \frac{3}{2e^2z^2}\left(\frac{k^3T^3}{\pi m}\right)^{\frac{1}{2}}.$$

Thus the mean free path is

$$L = \sqrt{\frac{3kT}{m}}\, t_c$$

$$= \frac{(3kT)^2}{8 \times 0.714\pi me^4z^4 \log \Lambda}$$

$$\approx \frac{1.8 \times 10^5 T^2}{N \log \Lambda},$$

TABLE 1.—*Mean Free Path of a Charged Particle for the Temperatures and Number Densities Shown*

Temperature (°K)	Mean Free Path of Charged Particles			
	$N = 10^8 \frac{\text{particles}}{\text{cm}^3}$	$N = 10^9 \frac{\text{particles}}{\text{cm}^3}$	$N = 10^{12} \frac{\text{particles}}{\text{cm}^3}$	$N = 10^{15} \frac{\text{particles}}{\text{cm}^3}$
10^3	1.4×10^4	19		
10^4	1.1×10^5	1.4×10^3	1.9	
10^5	$.91\times 10^8$	1.1×10^5	1.4×10^2	.19
10^8	$.79\times 10^{10}$	$.93\times 10^7$	1.1×10^4	14.5

TABLE 2.—*Variation of c/ω_p with Number Density*

N $\left(\frac{\text{Particles}}{\text{cm}^3}\right)$	c/ω_p
10^3	1.7×10^4
10^6	540
10^9	17
10^{12}	.54
10^{15}	1.7×10^{-2}

and is independent of the mass of the particle. In table 1 this quantity is given for $N = 10^6$ to 10^{15} particles/cm^3 and $T = 10^3$ to 10^6 °K. From this table it appears that the upper chromosphere is the region where the mean free path is measurable by optical methods. Near the photosphere the mean free path is too short and the thickness of the bright ring will probably be too small to be measured.

A radially outward moving current sheath moving with a velocity of the order of 1 km/sec produces a shock wave the thickness of which is of the order of

$$10\,\frac{c}{\omega_p} = 10\,\frac{3 \times 10^{10}\left(10^{-27}\right)^{\frac{1}{2}}}{\left(4 \times 10^{15} \times 25 \times 10^{-20}\right)^{\frac{1}{2}}} \approx \frac{1}{10}\ \text{cm}$$

(reference 13) which is also too small. Table 2 shows the variation of c/ω_p with density.

It is only in the upper chromosphere that the thickness of the shock is sufficient to be observable. Therefore, according to the present analysis, the bright ring should be observed to be expanding.

ACKNOWLEDGMENTS

The author would like to thank Drs. W. N. Hess, J. M. Burgers, and D. A. Tidman for fruitful discussions and Mr. J. Grunwald, and Mrs. E. Glover for help with programing the equations for the computer.

REFERENCES

1. GOLDBERG, L., and DYER, E. R., Jr., "The Sun," *In*: "Science and Space," (L. V. Berkner and H. Odishaw, eds.): 307–340, New York: McGraw-Hill, 1961.
2. DEJAGER, C., "Structure and Dynamics of the Solar Atmosphere," *In*: "Handbuch der Physik," v. 52 (S. Flugge, ed.): 80–362, Berlin: Springer-Verlag, 1959.
3. BIERMANN, L., "Der Gegenwärtige Stand der Theorie Konvektiver Sonnenmodell," *Vierteljahrsschrift Astron. Ges.*, **76**(4): 194–200, 1941.
4. THOMPSON, W. B., "Thermal Convection in a Magnetic Field," *Phil. Mag.* Ser. 7, **42**, 1417, 1951.
5. CHANDRASEKHAR, S., "On the Inhibition of Convection in a Magnetic Field," *Phil. Mag.* Ser. 7, **43**, 501, 1952.
6. DEJAGER, C., "The Development of a Solar Centre of Activity," *In*: "Vistas in Astronomy," vol. 4 (A. Beer, ed.): 143–183, Oxford, New York, Pergamon Press, 1961.
7. KIEPENHEUER, K. O., "Solar Activity," *In*: "The Sun," (G. P. Kuiper, ed.): 322–465, Chicago: University of Chicago Press, 1953.
8. THIESSEN, G., "The Magnetic Field Strength in Sunpots," *Naturwissenschaften*, **40**(7): 218–219, 1953.
9. MATTIG, W., "Die Radiale Verteilung der Magnetischen Feldstaerke in Normalen Sonnenflocken," *Zeits. Astrophysik*, **44**(4): 280–300, 1958.
10. COWLING, T. G., "The Growth and Decay of the Sunspot Magnetic Field," *Mon. Not. Roy. Astron. Soc.*, **106**(3): 218–224, 1946.
11. ROSENBLUTH, M., GARWIN, R., and ROSENBLUTH, A., "Infinite Conductivity Theory of the Pinch," Los Alamos Scientific Lab., New Mexico Rept. No. LA 1850, Sept. 14, 1954 (declassified 1958).
12. MINNEART, M., "The Photosphere," *In*: "The Sun," (G. P. Kuiper, ed.): 88–185, Chicago, University of Chicago Press, 1953.
13. ADLOM, J. A., and ALLEN, J. E., "The Structure of Strong Collision Free Hydromagnetic Waves," *Phil. Mag.*, **3**, 448, 1958.

SCIENTIFIC RESULTS OF OSO-1

JOHN C. LINDSAY

Goddard Space Flight Center

THE OSO-1 SPACECRAFT

A new era in the study of the sun from above the earth's atmosphere began with the successful launching of the first Orbiting Solar Observatory on March 7, 1962. Prior to this launch, measurements of solar radiations that are absorbed in the atmosphere had been made from balloons, rockets, and rather simple earth satellites. As a result of the various restrictions imposed by these techniques it was not possible to achieve a reasonable degree of observation continuity such as has now been accomplished by the OSO–1. The characteristics of the spacecraft that made this performance possible were the ability of the spacecraft to point instruments accurately at the sun, a circular orbit of approximately 575-km altitude below most of the radiation belt but above the atmosphere, and a tape recorder data storage system that supplied complete orbit coverage.

The OSO–1, like many other satellites, used the gyroscopic properties of a spinning body for stability. However, unlike other spacecraft, an integral part of the OSO–1 was a unique biaxial attitude control system to point instruments at the center of the sun. The spacecraft, figure 1, (Dolder, Bartoe, Mercure, Gablehouse, and Lindsay[1]) consisted of two parts, a wheel section and a platform upon which the solar cell array was mounted. The wheel section was spun to provide spacecraft stability, the spin axis being maintained approximately perpendicular to the solar vector with cold gas precession jets. Driving against the wheel section, an electrical servo-system oriented the platform perpendicular to the solar vector around the spin or azimuth axis. A second servo-system within the platform oriented an instrumentation section toward the sun in elevation. In this manner, the pointed instrumentation section was oriented along the solar vector in two axes. The spin of the wheel system was maintained by a gas system mounted on the end of arms.

A total of over a year's operation was obtained from the satellite with near perfect performance for approximately 3 months and real-time data recovery for a period of about 9 months. Although

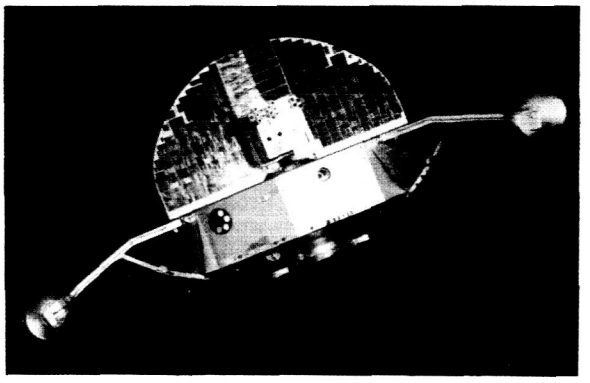


FIGURE 1.—OSO–1 spacecraft.

there were some problems with the pointing system during the life of the satellite, the long term pointing accuracy was within approximately 2 minutes of arc in both azimuth and elevation.

One might summarize the engineering firsts of OSO–1 as follows:

1. The first of the observatory spacecrafts, with demonstrated solar pointing more accurate than previously obtained with satellite or rocket systems.

2. First long-time operation of d.c. torque motors, bearings, slip rings, and other moving parts in the space environment for an extended

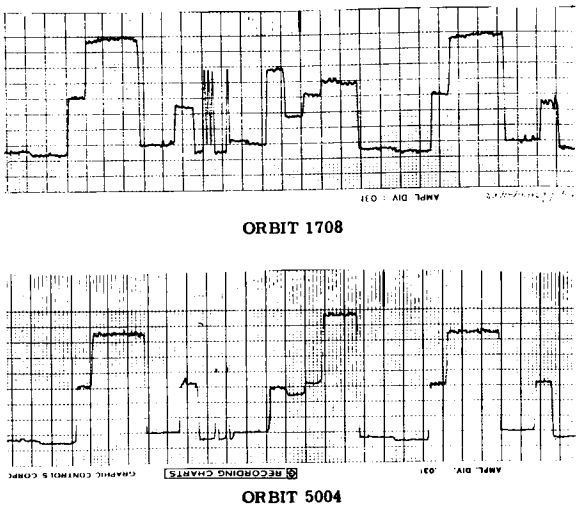


FIGURE 2.—Slip ring performance. (Data shown for Orbits 1708 and 5004 were transmitted through slip rings from sail to wheel section of the satellite.)

period of time, 18 months. An example of slip ring performance is shown in figure 2.

3. A unique damper to minimize spacecraft wobble.

4. First scientific spacecraft utilizing tape recorders to obtain complete orbit data coverage.

THE OSO-1 SCIENTIFIC EXPERIMENTS

Extreme UV Spectrometer

The primary solar oriented experiment aboard OSO–1 was a scanning spectrometer used for observation of the solar extreme UV radiation from 50–400A (Behring, Neupert, and Lindsay[2]).

During operation the spectrometer was pointed at the center of the solar disk within approximately 2 min of arc. In this orientation, radiation from the entire solar disk and inner corona passed directly through the entrance slit and struck a concave grating mounted in grazing incidence, the angle of incidence being 88°. The grating, an original ruled in a special glass by the Nobel Institute in Stockholm, had 576 lines per millimeter on a blank of 1 m radius of curvature. The exit slit and detector were mounted on a carriage which was driven on a circular rail so that the exit slit scanned along the Rowland Circle, where the spectrum was focused, from 10–400A. The plane of the exit slit was approximately perpendicular to the diffracted ray at all positions along the track, thereby keeping the spectral passband nearly constant for all angles of diffraction. The 50-micron entrance and exit slits provided a spectral passband of 1.7A and permitted resolution of lines 0.85A apart. The detector was a windowless photomultiplier developed by the Bendix Corp. specifically for use in this spectrometer. A tungsten photocathode was chosen to minimize response to wavelengths above 1500A, and to reduce changes in sensitivity due to variations of the emission properties of the cathode.

The spectrum obtained from the spectrometer over wavelengths of 170A to 400A is shown in figure 3. The brightest emission line in the region from 170A to 340A is the Lyman-alpha line of ionized helium at 304A. In addition, numerous other emission lines appear with combined flux comparable to, or somewhat greater than, that of the helium line. Resonance lines of heavy ions (Mg through Fe) are expected in this region, leading to attempts (Zirin, Hall, and Hinteregger[3]; Neupert and Behring[4]), to identify the more prominent features of the spectrum in terms of such lines. Other than the 304A line of He II Lyman-palha, the only lines identified with relative certainty are the 284A line of Fe XV and the 335A line of Fe XVI.

The months of March and April of 1962, were ideal for a study of the solar EUV spectra (Neupert, Behring, and Lindsay[5]) in that observations could be made on both a quiescent and a disturbed solar atmosphere. During the second week in March the sun was especially quiet, the sunspot number being zero on March 11. As the month

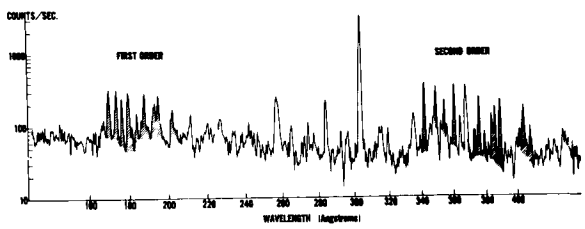


FIGURE 3.—Typical solar spectrum 170 A to 400 A.

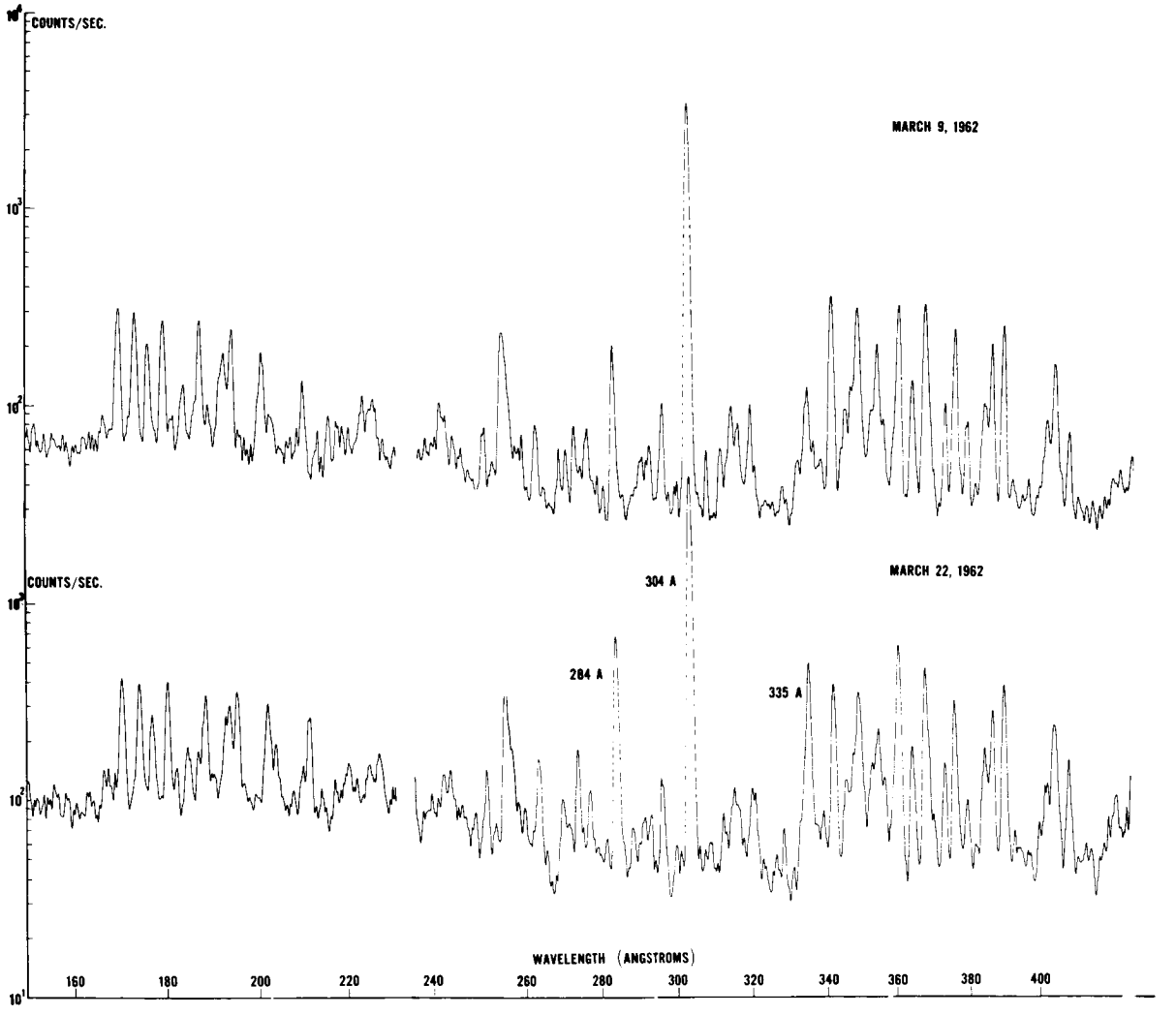


FIGURE 4.—Comparison of two spectra representing "quiet" and "active" sun.

progressed the solar rotation carried several centers of activity across the visible hemisphere of the sun. Definite enhancements in the solar spectrum were associated with these centers of activity.

Figure 4 presents two scans of the EUV spectrum which were obtained with a separation in time of approximately 10 days. During the first of these observations only one small region of activity was present on the solar disk. In spite of this low level of activity it is observed that the Fe XV and Fe XVI lines persist as two of the more prominent features of the spectrum. The second spectrum was obtained while several large and well-developed centers of activity were present on the disk. Comparing these two spectra we observe that the emission lines have increased in intensity but not all by the same amount. The Fe XV and Fe XVI lines, already prominent even in the absence of solar activity, have increased in intensity appreciably more than any other line observed with certainty in this spectral range. The He II line has also increased, but by a lesser amount.

The relationship of these observed counting rates to several ground-based measurements of solar activity is presented in figures 5 and 6. In figure 5, the He II radiation is compared with daily values of the solar flux at 2800 mc, and with the Zurich Provisional Relative Sunspot Number

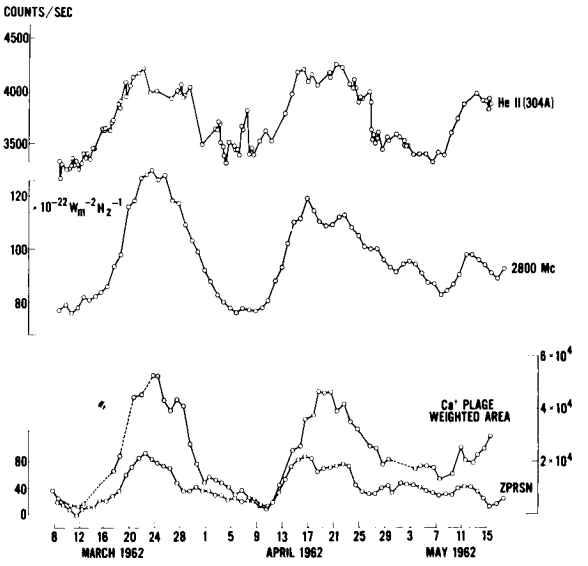


FIGURE 5.—He II Lyman-alpha line compared with ground-based measurements of solar activity.

(ZPRSN). Also shown is an estimate of the calcium plage area, each area being weighted by the estimated intensity of the area on a scale from 1 to 5. Values for this computation were supplied by the McMath-Hulbert Observatory. In figure 6, the daily values of solar flux at 2800 mc and the Zurich Provisional Relative Sunspot Number are compared with radiation due to the coronal lines of Fe XV (284A) and Fe XVI (335A). The estimated calcium plage intensity is also shown.

The spectral lines chosen for presentation here were selected because they are reliably identified

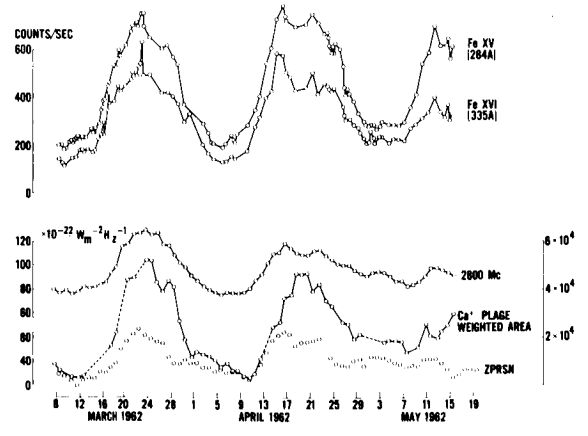


FIGURE 6.—Fe XV and Fe XVI lines compared with ground-based measurements of solar activity.

with particular ions, not because they convey more than any other line in the spectrum the changes in solar flux which occur with the appearance of plage areas. In terms of fractional changes in intensity, these three lines represent the extremes which have thus far been observed in the spectral region from 171A to 400A; only a few faint lines have smaller nonflare variations than the He II Lyman-alpha line, while no other lines have increases as great as those observed for 284A and 335A. A summary of the average increases in counting rates for the period from March 9 to March 23, 1962, a period of increasing solar

TABLE I.—*Increases in Solar EUV Spectrophotometer Counting Rates March 9, 1962 to March 23, 1962*

Spectral range	Average counting rate increase
	Percent
171A–228A	55.
229A–300A	80.
229A–305A	52.
305A–400A	50 (estimated).

activity, is given in table 1. The increase, weighted by the intensity of each line, is computed for the range from 171A to 305A, using 60 reliably observed lines. The increase in the range from 305A to 400A can only be estimated because of the masking effect of second order images above 342A. The values given in table 1 are, of course, appropriate only for the particular interval in time for which they were computed.

The initial analysis of only three lines (He II 304A, Fe XV 284A, Fe XVI 335A) already indicates that the relative prominence of spectral lines may depend upon the age of the center of activity which is responsible for the increased radiation. As an example of this, one may observe (fig. 6) that the maximum emission in the Fe XV apparently occurs later in time than the maximum for the 2800-mc radio flux or for the plage areas observed during March, April, and May. In addition to such a slowly changing effect, one may note that localized perturbations appear (March 7–9 and April 16–17) for which the relative increases are considerably different for

the helium and the iron lines. It appears that in these instances we are observing phenomena localized at particular levels in the solar atmosphere.

1–11A X-ray Experiment

An experiment to monitor solar X-ray was flown on OSO–1 by W. A. White[6] and R. M. Young of Goddard Space Flight Center. The detector was a beryllium window ion chamber with a Xenon filling. The conversion efficiency as a function of wavelength is shown in figure 7.

The full-scale sensitivity is dependent upon the shape of the input spectrum. If the spectral shape obtained by Pounds, Willmore, et al.[7] from the satellite Ariel, which is consistent with a 2.8 $(10)^6$ °K plasma, is assumed the wavelengths contributing to the output current are in the interval 3–11A and the full-scale sensitivity is $1.8(10)^{-3}$ ergs cm^{-2} sec^{-1}. For comparison with earlier measurements by R. W. Kreplin, et al.[8] over bandwidths specified as 2–8A, the full-scale sensitivity of the OSO–1 experiment is $3.6(10)^{-4}$ ergs cm^{-2} sec^{-1}.

If one computes the continuum flux to be expected at these wavelengths from the entire solar corona, the theoretical flux following the work of Elwert[9] for an isothermal corona with

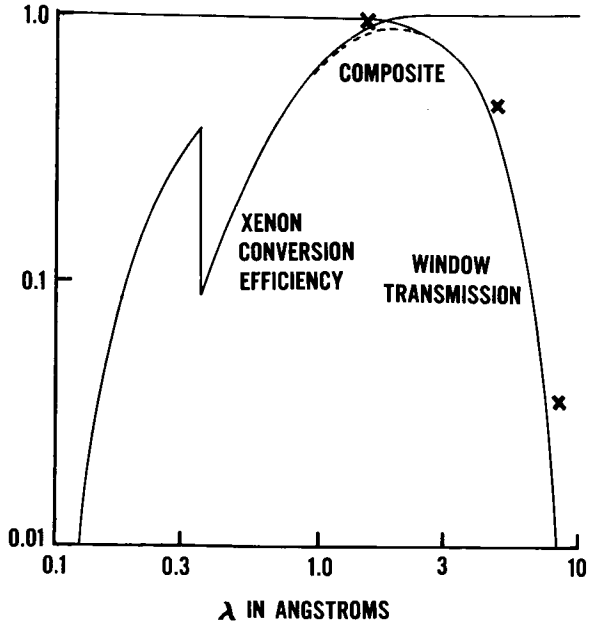


FIGURE 7.—OSO–1 ion chamber efficiency as a function of wavelength.

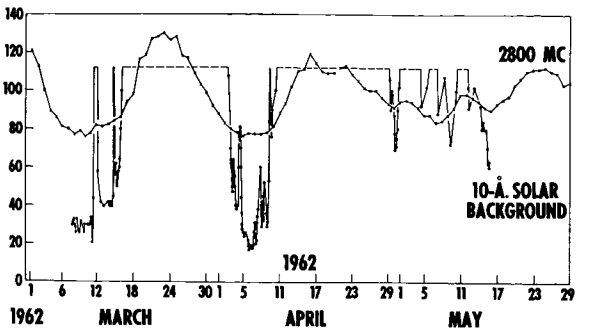


FIGURE 8.—Slowly-varying component of solar X-rays compared with 2800-mc radio measurements.

Allen's[10] electron density profile at $2.8(10)^6$ °K falls short of the lowest value measured by OSO–1 by a factor of the order of 15. In fact, the entire corona would have to be at a temperature in excess of $3.5(10)^6$ °K to meet the lowest OSO–1 flux using such an all-continuum model. If one assumes the contribution from line emission in excess of the flux from continuum emission by a factor of 15, the corona in its entirety would have to be at a temperature of about $2.4(10)^6$ °K; to explain the lowest flux measured by OSO–1. For more than 50% of the time the OSO–1 flux exceeded this lowest value by at least a factor of 10.

From the fact that most of the time the solar X-ray flux measured by OSO–1 was much larger than the lowest value (which is already uncomfortably high for an isothermal corona with uniform density profile, even with appreciable line emission), we can only conclude that the source of X-rays of less than 11A wavelengths must conform to localized regions such as plages, as has been observed for longer wavelengths.

A comparison of the slowly varying part of the 10-Angstrom X-ray flux with 2800-mc radiation confirms that the localized sources of solar X-rays are associated with centers of activity such as Ca plages and/or sunspot groups. Figure 8 shows the time history of both fluxes for about 2.5 solar rotations in the early life of OSO–1. It can be seen that the smoothed X-ray flux correlates with the excess 2800-mc flux above a background of about 75 flux units appropriate for the "quiet" sun at that phase of the solar cycle (Covington and Harvey[11]).

The lowest X-ray flux measured by OSO–1 (April 6, 1962) was: for $\lambda < 8$A, $3.6(10)^{-5}$ erg cm^{-2}

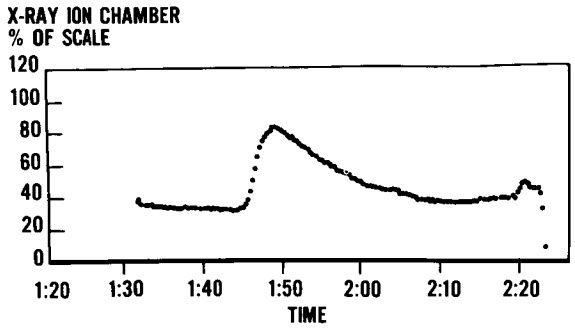


FIGURE 9.—Small X-ray flares.

sec^{-1}; for $\lambda < 11A$, $1.8(10)^{-4}$ ergs cm^{-2} sec^{-1}. This may be considered an upper bound on the X-ray flux from the "quiet" sun. This flux occurred at a time when only 3 plages of area $\geq$ 1000 millionths of a solar hemisphere were visible on the disk. The nearest plage behind the west limb had set 3 days previously, and the nearest behind the east limb was not to rise until 1.5 days later. If we assume that the X-ray emission is coming from 3 "pillbox" volumes, one associated with each of the 3 plages, each having its base area equal to the plage area and furthermore having a height equal to $\frac{1}{2}$ the diameter of its base, the source volume is 3.85×10^{29}cm.[3] The volume above on of the plages, No. 6379, was observed by Billings[12] on April 9, 1962, to have a faint continuum enhancement from which he estimated the electron density $n_e = 0.5(10)^{10}$. Using this as the electron density, and assuming (as is most likely from such a weak source, Hallam and Young,[13]) the electron temperature was not greater than $3.0(10)^6$ °K, one finds that in order to obtain the X-ray flux measured on OSO–1 the ratio of line emission to continuum emission must be at least 10:1.* If T_e was no greater than the $2.8(10)^6$ degree value obtained from the Ariel spectrum of April 27, the ratio of line emission to continuum emission must have been around 30:1.

Up to now we have discussed only the slowly varying component of the X-ray emission: In addition to these quasi steady-state conditions, transient events (X-ray flares) lasting usually from 10 min to a couple of hours were frequently observed. Such an event is shown in figure 9, and should be compared with the quiet period of similar duration shown in figure 10. The particular event of figure 3 contains a total energy below 11A of approximately $2(10)^{27}$ ergs.

During the 9-day interval between launch and March 16, 1962 (at which time the rising of plage No. 6370 on the east limb supplied enough X-ray emission to carry the experiment off-scale), approximately 60 X-ray flare events lasting from 10 min to 1 hr were seen, and 4 events were seen to last about 5 hrs.

During this same interval (1620 U.T. March 7, 1962, to 1620 U.T. March 16, 1962) some 33 H–α flares were reported by ground-based observatories. Of these H-α flares, 6 would have been unobservable from OSO–1 for various reasons (satellite night, failure to command data storage readout, etc.). Of the remaining 27 H-α flares, 3 occurred while the X-ray experiment was still off-scale because of a previous large event. This leaves 24 H–α flares which can be tested for correlation with the X-ray flares. Of this group of 24, it appears that 11 correlate well, 3 definitely have no counterpart in X-rays, and the remaining 10 are doubtful because of insufficient data or an excessive time difference (>10 min). Conversely, there are 6 full-scale or greater X-ray events for which no H-α flare was reported even though observations were presumably being made at the time. Certainly more observations will be required before a definite statement can be made regarding a correlation or lack thereof between H–α flares and X-ray flares.

In looking for correlations with Sudden Ionospheric Disturbances, all X-ray events exceeding the full-scale saturation level were barely detectable (if observing conditions permitted) in Sudden Phase Anomaly data for VLF trans-

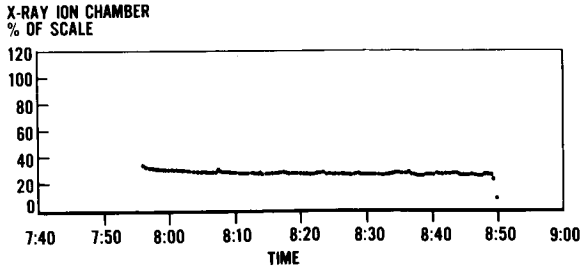


FIGURE 10.—Typical "quiet" X-ray period.

*A similar conclusion has been suggested by Pounds, Wilmore, Brown, Norman, and Sanford (reference 7) based upon the Ariel data.

missions via the *D* layer. Only the large event of March 13 was seen in ionospheric indices other than SPA's. Correlation with transients in the 2800-mc solar flux is good; but again, full-scale X-ray events are represented by extremely small events (1 to 2 flux units) in the 2800-mc data.

Several apparent associations of certain X-ray flares into groups displaying a definite pattern were observed (White[6]). Figure 5 shows such a grouping. Similar groupings are present in the data for the first week in April; in fact, the one particular March group shown in figure 11 has an exact April counterpart 27.1 days later, with identical time-separations between events and with identical peak excursions above mean background level. The envelope joining the peaks of the flares within a group is found:

(a) to be a straight line, and
(b) to have the same slope (with either positive or negative sign) from group to group.

These characteristics of the flare groups indicate a constant time rate-of-change of X-ray source strength. Source strength is a function of electron density, of temperature, and of volume. It is difficult to see why any time variation of either electron density or temperature would be of such a particular nonlinear nature as to constrain the source strength to vary linearly with time. One is left with the concept of a volume which is either growing or diminishing at a constant rate, and which on occasion serves as a reservoir of high-temperature electrons and ions interacting to produce the X-ray flares.

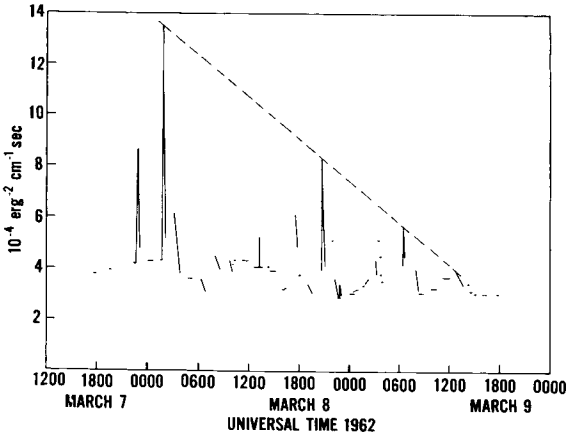


FIGURE 11.—Grouping of X-ray flares for period March 7–March 9, 1962.

Hydrogen Lyman-alpha Measurements

The solar hydrogen Lyman-alpha flux was monitored from OSO–1 with an experiment performed by Kenneth Hallam and Robert Young[13] of the Goddard Space Flight Center. The detector used was a CS_2 filled ion chamber with a L_iF window.

The average solar flux measured in the spectral region 1050A to 1230A was 4.9±0.1 ergs/cm²-sec for the first 40 orbits. Lyman-alpha contributes about 95% of the flux in this bandpass. The remainder is mainly due to Si III, λ 1206.5A. After the first 230 orbits, the ion chamber sensitivity declined at about 20% per week. This, however, had no effect on short term relative measurements.

A flare of importance 2+ on March 13 caused a peak enhancement in Lyman-alpha of 5.3%. An increase of 6.8% was observed during an importance 3 flare on March 22. This represents a local brightening in Lyman-alpha of between 5 and 150 times, depending on the background situation.

20 Kev-1 Mev Gamma-ray Experiments

Two experiments to search for solar gamma-radiation in the 20 Kev to 1 Mev region were flown by K. J. Frost, E. D. Rothe, K. L. Hallam, W. A. White, and H. M. Horstman[14] of the Goddard Space Flight Center. The first experiment consisted of a thin NaI(T1) crystal (2.22 cm diameter×0.15 cm height) scintillation counter surrounded by a copper shield. The objective of this experiment was to search for solar bremsstrahlung bursts in the 20 to 100 Kev range. The second experiment consisted of three scintillation counterdetectors. Two NaI(T1) (3.8 cm diameter×3.8 cm height) crystals were placed in the wheel section of the observatory and one (3.8 cm diameter×5.08 cm height) CsI(T1) crystal was placed in the pointed section. These crystals measured the spectrum of gamma rays between 0.100 and 1.00 Mev with particular emphasis on the 0.511-Mev positron-electron annihilation line and its temporal variation.

The results of the 20–100 Kev experiment indicate that the flux from the quiet sun cannot be in excess of 3.40 ± 0.95 photons/cm²-sec in this energy range. At this time no solar bremsstrahlung bursts have been detected. The data survey thus far includes most of the information acquired during the months of March and April 1962. It is pertinent to note that no solar proton events occurred during this period, at the time observations were being made. The counting rate in the 0.1 to 1 Mev range was found to be 4.7 ± 0.5 counts/cm²-sec. The data also indicate that the upper limit of the positron-electron annihilation radiation flux from the sun is 0.6 ± 0.2 photons/cm²-sec.

Another gamma-ray experiment was flown by Laurence E. Peterson,[15] University of California, La Jolla, Calif. His experiment was designed to search for extra terrestrial gamma rays in the 50 Kev–3 Mev energy range. The apparatus consisted of a 2.5 cm diameter $\times$ 1.25 cm height NaI counter with a 0.5 cm lead collimating shield for 50–150 Kev photons. A Compton telescope, consisting of a 3.2 $\times$ 3.2 cm NaI counter in coincidence with a phoswich type NaI counter (5.1 cm diameter $\times$ 5.7 cm height) provided a directional detector for gamma rays between 0.3 and 3.0 Mev. The counting rate of the low energy telescope near 0 geomagnetic latitude was 1.0 ± 0.1 counts/cm²-sec, most of which was local cosmic ray produced background. The solar flux at the earth was less than 1 photon/cm²-sec or 1.6×10^{-7} ergs/cm²-sec between 50 and 150 Kev. No increases were noted during the flares on March 11 and March 22, 1962, and no significant variations above background over the celestial sphere have been observed. Typical total rates measured by the phoswich NaI counter at 0 geomagnetic latitude were 0.40 and 0.18 counts/cm²-sec for photons with energy losses of 0.3–1.0 and 1.0–3.0 Mev, and 0.35 counts/cm²-sec for particles losing more than 1 Mev. The respective rates at 40° geomagnetic latitude were 0.64, 0.44, and 1.6 counts/cm²-sec. Most of the gamma rays in this energy region, immediately above the earth's surface, were either due to earth albedo or were of local origin.

Proton-Electron Experiment

A proton-electron analyser was flown on OSO-1 by Carlton D. Schrader, Aerospace Corp., and R. C. Kaifer, J. A. Waggoner, J. H. Zenger, and S. D. Bloom[16] of the University of California, Lawrence Radiation Laboratory. The objective of the experiment was to determine the time and position variations of the fluxes of protons of energies greater than 1.5 Mev and electrons of energies greater than 110 Kev near the lower boundary of the inner Van Allen belt. The detector utilized the principle that in certain scintillators protons and electrons produce fluorescent pulses of distinctly different decay times. This made it possible, through electronic pulse shape discrimination, to employ a single crystal on asingle photomultiplier to detect and separately count both protons and electrons.

One of the most interesting preliminary results was the discovery of a number of "warm spots" (as contrasted to the anomalous south Atlantic radiation "hot spot") where the electron flux is more intense, by factors up to 50, than the average intensity. These warm spots are apparently constant in time, intensity, and position. They are all located between latitudes 33°N and 33° S (the limiting orbital latitudes of the satellite) and occur over Madagascar, western Australia, eastern Australia, northwest of Hawaii, off lower California and in the south Pacific. Preliminary analysis has included the plotting of the warm spot intensities as a function of the natural trapped particle coordinates, B and L. The warm spot intensity plots form well-defined curves in B–L space, which differ, however, from curves for similar B and L plotted from data recorded over other geographical locations. Thus, instead of a previously supposed longitudinal invariance, these data seem to indicate that at this altitude (near 575 km) there is a definite dependence on longitude. The experimenters are attempting, in a preliminary theoretical interpretation, to explain the warm spots as being due to trapped radiation which is being supplied at equilibrium rates from the lower Van Allen zone. As these low-altitude trapped electrons drift eastward across the Americas and the Atlantic they are lost because of the South Atlantic anomaly which causes them to mirrow below sea level. This accounts for the low intensities for a given B and L in this region. Even though most of the electrons have just been wiped out in their passage across the Atlantic, warm spots are already reestablished just east of

Africa. This is regarded as evidence that the warm spots are being constantly supplied from higher altitudes.

The general proton data exhibited less structure, but an apparent small monotonic increase of proton intensity as a function of time is being investigated. To check this point further, it is planned to continue to record additional data (available from the satellite on a real time basis until July 1962). Only one very low intensity proton warm spot has been observed so far, and this is above the Indian Ocean, between Africa and Australia.

Also of considerable interest were the many narrow, but intense, peaks of both protons and electrons which frequently occurred superimposed on the "normal" structure, presumed to be due to precipitation of previously trapped particles by magnetic disturbances. Many such intense peaks were found; e.g., in the data of March 12, 13,and 14, 1962, a period of unusual solar activity. The artificial radiation belt formed by the Starfish high-altitude nuclear test (July 9, 1962) was also observed over a limited area and the subsequent time history recorded.

High Energy Gamma Ray Experiment

The high energy gamma ray detector aboard the OSO–1 was designed by G. G. Fazio[17] of the University of Rochester and the Smithsonian Astrophysical Observatory to provide the first view of a solar flare in the >100 Mev region of the electromagnetic spectrum. A partial analysis of the data has shown no evidence for this radiation from the sun, even during the importance 3 flare of March 22, 1962, and the importance 2+ flare of March 13, 1962. The upper limit of the flux from each of these flares was estimated to be 10^{-2} photons/cm^2-sec, and that for the quiet sun, 7×10^{-3} photons/cm^2-sec. The flux from the total sky was less than 6×10^{-3} photons per/cm^2-sec steradian. Locally produced background radiation limited the sensitivity of the detectors. A more complete understanding of this background radiation will permit a sensitivity of 10^{-3} photons per cm^2 per sec from the sun.

Solar Neutron Experiment

An experiment to detect solar neutrons was flown on OSO–1 by W. H. Hess,[18] of the University of California, Lawrence Radiation Laboratory and Goddard Space Flight Center. The detector was a $B^{10}F_3$ proportional counter and the purpose was to observe diurnal variations of its counting rate, as well as the Hames sunset effect. Combining data from 42 orbits gave essentially no presunset maximum to within 10%. However, there appeared to be a variation of the neutron count rate from day to night. The preliminary ratio of count rates day/night is 1.08 ± 0.01. It is currently uncertain whether there is associated with this daytime neutron excess a concurrent proton excess.

SUMMARY OF OSO-1 OBSERVATIONS

Observations of the sun were made by OSO–1 for a time period corresponding to approximately three solar rotations on almost continuous basis and on a real time over the ground stations basis for approximately 1 year. The results reported here were based upon partial analysis of the first 3 months data.

The observations of the solar spectrum between 170A and 340A have shown:

1. That the He II (304A) emission is enhanced by a factor of about 33% during a period when the Zurich Provisional Relative Sunspot Number increased from zero to a maximum of 94 and the 2800-mc flux varied from approximately 76 to 125×10^{-22}wm^{-2}Hz^{-1}.

2. The Fe XV (284A) and Fe XVI (335A) coronal lines were enhanced during the same period by a factor of approximately four.

3. The enhancement of He II (304A) and Fe XV (284A) and Fe XVI (335A) due to plage activity was larger than enhancements due to flares that occurred during the 3-month interval of the observations.

4. The variations in intensity of the He II (304A), Fe XV (284A) and Fe XVI (335A) represent the extremes observed. If one averages 60 of the reliably observed lines between 171A and 342A the enhancement is between 50% and 80% for the time interval March 9 to March 23.

5. Although there appears to be a gross correlation between solar activity indices (such as 2800-mc flux) and the He II, Fe XV, and Fe XVI fluxes, there are indications that the relative prominence of the spectral lines may depend upon the age of the center of activity.

The 1–11AX-ray measurements have observed a slowly varying component which correlates with the slowly varying component of the 2800-mc solar radiation. A model for these quasi-stable X-ray sources which fits the OSO–1 data postulates localized sources having the same horizontal extent as Ca plages with thicknesses proportional to their diameter, and having an electron temperature of $2.8(10)^6$ °K or greater and an electron density of about $5(10)^9$ electrons per cm^3. For these conditions it is also necessary that the ratio of line emission to continuum emission be at least 10:1 and more probably 30:1.

In addition to a slowly varying component, transient events (X-ray flares) lasting from 10 min to a few hours were frequently observed. Correlation has been attempted with $H-\alpha$ flares with somewhat ambiguous results. X-ray flares were frequently observed to be associated in groups possessing a characteristic pattern; the concept of a source volume varying linearly with time is suggested to account for the linear envelope of a flare group.

The average hydrogen Lyman-alpha flux was approximately 4.9 ergs/cm^2-sec and for the first time an enhancement of Lyman-alpha has been observed during solar flares. The enhancement was 5.3% during a class 2+ flare and 6.8% during a rather unusual class 3 flare (practically no geophysical effects were observed for this flare).

No solar bremsstrahlung X-ray bursts (10–100 Kev) have been found in the data examined to date. An upper limit for "quiet" sun X-rays of this energy range was found to be 3.40 ± 0.95 photons/cm^2-sec. For 50–150 Kev X-rays the upper limit found was 1 photon/cm^2-sec. No gamma-rays >100 Kev have been observed.

The proton-electron experiment discovered a number of "warm spots" below the normal inner belt. These radiation "warm spots" apparently were constant in time, intensity, and position. They were found over Madagascar, western Australia, eastern Australia, northwest of Hawaii, off lower California and in the south Pacific.

To date no high energy gamma-rays (Energies >100 Mev) have been detected from the sun either during a "quiet" sun or a solar flare.

The solar neutron experiment has detected a very slight excess in neutron count from day to night, 1.08 ± 0.01. It is currently uncertain whether this excess is real and whether there is a corresponding proton excess.

REFERENCES

1. Dolder, F. P., Bartoe, O. E., Mercure, R. C., Jr., Gablehouse, R. H., and Lindsay, J. C., Space Research III, COSPAR, Preiester Ed., North Holland Publishing Co. (1963).
2. Behring, W. E., Neupert, W. M., and Lindsay, J. C., Space Research III, COSPAR, Preiester Ed., North Holland Publishing Co. (1963).
3. Zirin, H., Hall, L. A., and Hinteregger, H. E., Space Research III, Preiester Ed., North Holland Publishing Co. (1963).
4. Neupert, W. M., and Behring, W. E., J. Quant. Spectrosc. Radiat. Transfer, vol. 2, 527 (1962).
5. Neupert, W. M., Behring, W. E., and Lindsay, J. C., Space Research IV (to be published), 1964.
6. White, W. A., Space Research IV (to be published), 1964.
7. Pounds, K. A., Willmore, A. P., Bowen, P. J., Norman, K., and Sanford, P. W., Proc. Roy. Soc. (in press).
8. Kreplin, R. W., Chubb, T. A., and Friedman, H., Jour. Geophys. Res., **67,** 2231 (1962).
9. Elwert, G., Jour. Geophys. Res., **66,** 391 (1961).
10. Allen, C. W., Monthly Notices Roy. Astr. Soc., **106,** 22 (1946).
11. Covington, A. E., and Harvey, G. A., Astrophys. Jour. 132.
12. Billings, D. E., private communication.
13. Hallam, K. L., and Young, R., presented at spring meeting of AGU, April 1963.
14. Frost, K. J., Rothe, E. D., Hallam, K. L., White, W. A., and Horstman, H. M., presented at spring meeting of AGU (Apr. 1963).
15. Peterson, L. E., presented at spring meeting of AGU, April 1963.
16. Schrader, C. D., Kaifer, R. C., Waggoner, J. A., Zenger, J. H., and Bloom, S. D., presented at spring meeting of AGU, April 1963.
17. Fazio, G. G., presented at spring meeting of AGU, April 1963.
18. Hess, W. H., presented at spring meeting of AGU, April 1963.

EMISSION OF EXTREME ULTRAVIOLET RADIATION FROM SOLAR CENTERS OF ACTIVITY

WERNER M. NEUPERT

Goddard Space Flight Center

Line emissions at 284A, 304A, and 335A in the solar extreme ultraviolet spectrum observed by OSO–1 have been compared with 2800 Mc emission and other indicators of solar activity. Increases in these ultraviolet fluxes can be localized to active regions on the solar disk. Changes in the intensities of the lines are discussed in terms of changes in electron temperature and electron density of the active regions. Certain assumptions made in the calculation of a coronal electron temperature from the relative populations in two stages of ionization of an element are examined, with the conclusion that this calculation may not be valid for the particular case of the solar corona, yielding a fictitiously low value if there is an outward expansion of ions from the base of the corona into a region of higher electron temperature. A nonequilibrium formulation for the distribution of ions is presented which may better describe the actual ion distribution in the corona. An approximate calculation shows that an average radial velocity, outward, and small compared with the average thermal velocity of the ions, is sufficient to explain the observed increase in height of the maxima of emission of the Fe ion sequence with increasing stages of ionization in the undisturbed corona.

INTRODUCTION

A grazing incidence spectrometer, operating as a pointed experiment on OSO–1 (1962 ζ1), has provided the first observations of variations of solar extreme ultraviolet (EUV) fluxes due to the presence of solar centers of activity on the sun's surface. In another paper presented at this conference,* these variations have been compared with fluctuations observed in other ranges of the solar electromagnetic spectrum; in this paper, the variations in EUV fluxes will be discussed in terms of changes in coronal electron densities and coronal temperatures that may be derived from them. Since the spectrometer has no inherent spatial resolution, this discussion will center around a comparison of EUV fluxes from the entire solar disk, when no disturbed regions are present, with fluxes observed when a well defined center of activity is present on the disk. The discussion is further restricted to three of the more reliably identified lines in the spectral range of the instrument: 284A (Fe XV), 304A (He II), and 335A (Fe XVI).

*"The Solar Spectrum from 50A to 400A," pp. 596–601.

VARIATIONS IN SELECTED EUV EMISSION LINES

A summary of the nonflare variations observed in the above mentioned lines is given in figures 1 and 2, in which each datum point represents the average of one orbit's observations (six to eight spectral scans). The error bars indicate one standard deviation in the data. The relationship which these observed counting rates have to several ground-based measurements of solar activity is also presented. In figure 1, the He II Lyman-alpha radiation is compared with daily values of the solar flux at 2800 Mc (National Research Council, Ottawa, Canada) and with the Zurich Provisional Relative Sunspot Number (ZPRSN). Also shown is an estimate of the calcium plage intensity, computed for each day by summing over the observed (foreshortened) area of all plages, each area being weighted by the estimated intensity of the plage, on a scale from 1 to 5. Areas were taken in millionths of the solar hemisphere. Values for this computation were supplied by the McMath–Hulbert Observatory.

A similar presentation is made in figure 2 for observations of the coronal lines of Fe XV and

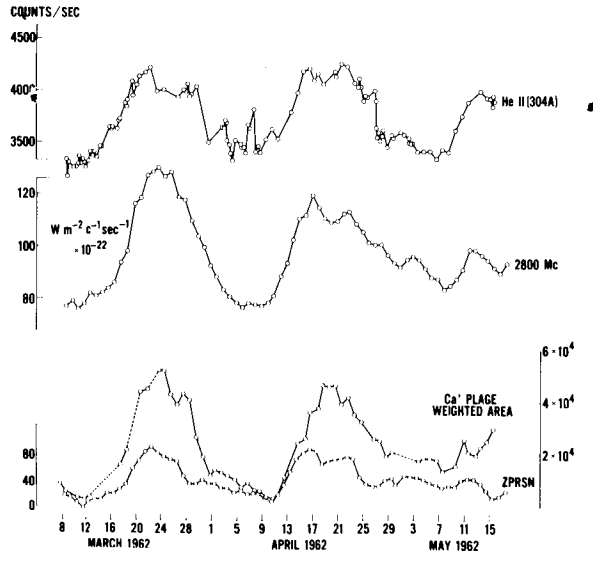


FIGURE 1.—Comparison of the He II (304A) line with measurements made at radio and optical wavelengths.

Fe XVI. Two methods were applied for estimating the instrument sensitivity at 284A and 335A. In the first of these, solar fluxes observed at those wavelengths by Hall, Damon, and Hinteregger,[1] whose instrument had been calibrated, were compared to counting rates obtained by the GSFC instrument on days of comparable solar activity. The 2800 Mc daily flux was used as an index of this activity. In the second, the sensitivity was computed from estimated values of the grating reflectivity and detector efficiency. Sensitivities of 4.2×10^5 photons/cm²-sec-count at 284A and 8×10^5 photons/cm²-sec-count at 335A, representing averages of the values given by the two methods, were adopted and satisfied the need for a relative calibration at the two wavelengths. Their use as absolute values of the sensitivity is not recommended because of the limited validity of the calibration procedure.

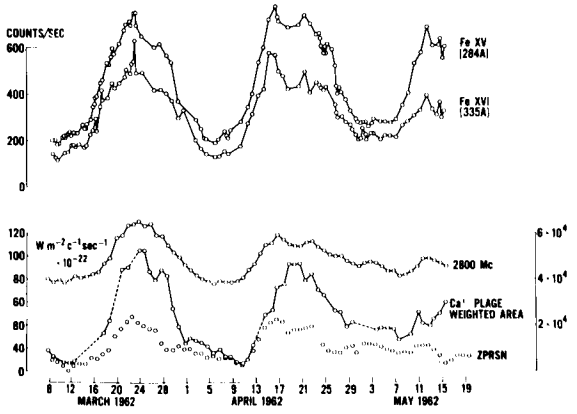


FIGURE 2.—Comparison of Fe XV (284A) and Fe XVI (335A) lines with measurements made at radio and optical wavelengths.

DISCUSSION OF THE DATA

Observation of the Quiet Corona

The use of OSO–1 as a long-term stable platform has permitted the acquisition of a solar EUV spectrum which can tentatively be associated with a corona not disturbed by visible centers of activity. This situation presented itself for a short period of time on March 11, when a ZPRSN of zero was reported. The American Relative Sunspot Number was reported as zero on both March 10 and March 11. In addition, no large active regions had been associated with the face of the sun turned toward the earth on March 11 during the preceding 6 months. A careful analysis of the more intense emission lines, made for the period from March 7 to April 5, demonstrates that the lowest counting rates of the period were indeed observed when the sunspot number was near zero and the calcium plage area on the sun was also at a minimum. However, it is also clear from an examination of figures 1 and 2 that no close correlation can be assumed to exist between the EUV fluxes and any of the ground-based observations. As figure 2 shows, counting rates from the iron lines were lower on 1 May, when the Sunspot Number was 49, than on 15 May, when it was at 13. Likewise, although the agreement between the radio data and EUV coronal fluxes at times is striking (as for instance in the interval from April 13 to April 25, a period of considerable solar activity), this similarity is not consistent, as the data taken from April 29 to May 15 shows: A small maximum observed at 2800 Mc on May 1 to May 5 does not appear at 284A whereas on the succeeding days, May 9 to May 15, a larger peak is observed at 284A than at radio wavelengths. The anomaly can be eliminated, at least in this particular instance, by observing that the flux at each wavelength depends on the age of the center of activity being observed. The region under

observation on May 1 to May 5 is a center of activity approximately one solar rotation old whereas the plage areas observed on May 9 to May 15 are remnants of active regions 4 to 5 rotations old. The trend in the data is consistent with a model for a center of activity (CA) described by De Jager,[2] in which the center is initially localized in the lower altitudes of the solar atmosphere and gradually expands into the quiet corona. After the spot phase of the CA has passed, the region remains as a magnetic bipolar region which may have quiescent filaments and perhaps coronal rays associated with it. The continued enhancement of the Fe XV and Fe XVI lines after all sunspots have vanished in the region may be an indication of remaining coronal rays. In any event, these observations suggest that it is necessary to have knowledge of the recent past history of solar activity as well as current data in order to make a correlation with the EUV radiation.

Having established an emission spectrum corresponding as nearly as the data permit to an undisturbed corona as observed approximately four years after sunspot maximum, we may compute an electron temperature for the quiet corona which is implied by the observed spectrum The reason for choosing an undisturbed spectrum. for this calculation is that we are forced to make an assumption of the spherical symmetry of the corona since the spectrometer has, inherently, no spatial resolution and this assumption is best satisfied when large scale structures, as evidenced by EUV enhancements, are absent An electron temperature may be derived[3] from the observed ratio of intensities of two spectral lines originating from two different ionic species of the same element. Such a calculation has validity only if both radiations originate in the same region of the corona at some well-defined electron temperature, and only if the ionization-recombination equilibrium assumed by the theory is actually present. These assumptions will be examined in the following sections.

The ionization formula used to relate the relative populations in adjacent states of ionization was that of Elwert.[4–5] Excitation cross-sections were computed from the expression derived by Woolley and Allen on the basis of the classical Thompson atom. The oscillator strengths were taken from Varsavsky.[6] It is found that an electron temperature of 1.75 million degrees in that part of the corona assumed to be the source of both the Fe XV and Fe XVI radiations best fits the observed data. The estimated possible error in this value arising from uncertainty in the relative sensitivity of the spectrometer at the two wavelengths is 100,000°K. This result contrasts sharply with the value of 800,000°K usually obtained from observation of the forbidden lines of Fe X and Fe XIV. Since both results are ultimately based on the same theory of ionization equilibrium, we are forced either to question the validity of the ionization equilibrium assumed or to postulate that a range of electron temperature exists in the corona. Indeed the very observation of six stages of ionization of iron, from Fe X through Fe XV, excepting Fe XII in the visible, coupled with the Fe XVI resonance transitions in the ultraviolet, leads to the same conclusion.

EUV Radiation from Solar Centers of Activity

A discussion of the sources of radiation from Fe XV and Fe XVI may be extended to include centers of activity and plages since it appears from the data that emission from these two ions is intimately associated with these disturbances. In this section we will consider whether or not a unique change in electron temperature or density or both can be obtained for the corona above plages from the EUV data. A tentative analysis has been made for changes observed in the 284A and 335A counting rates between March 11 (ZPRSN equal to zero) and March 25 (ZPWSN equal to 74). Although the latter date does not represent the period of greatest deviation of the spectrum from quiet sun conditions, it was chosen because of the relatively well-defined configuration of three plage areas near the central meridian at that time. During this period the Fe XVI radiation increased by a factor of 3.85, the Fe XV radiation by 3.37, and the Fe XIV radiation (as estimated from optical observations of λ 5303) by 2.50. It is obvious that the center of activity represents a more complex change than simply an increase in density in a volume having a uniform and non-changing electron temperature. On the other hand, it is also impossible to explain the changes in

emission as exclusively due to an increase in electron temperature in a homogeneous volume of the corona; that is, no one unque increase in electron temperature the center of activity above that of the quiet corona can explain the observed spectral variations. From the nearly constant Fe XVI/Fe XV ratio there is only a small temperature increase, 100,000°K, whereas the increase implied by the Fe XV/Fe XIV ratio (using λ 5303 data) is more than $2\frac{1}{2}$ times as large. The inability to obtain consistent results under the assumption of an isothermal center of activity again suggests that not all spectral lines are emitted uniformly from the corona but that those regions (or region) emitting Fe XV and Fe XVI radiation from the quiet corona become relatively more predominant over centers of activity and plages.

EFFECT OF A CORONAL EXPANSION ON THE CALCULATION OF AN ELECTRON TEMPERATURE

Discussion of the Ionization Equation

The inconsistencies discussed in the foregoing sections, as well as the long-standing discrepancy between the electron temperature of the quiet corona obtained from forbidden emission line intensity ratios and other values computed from the electron density gradient has led to an analysis of the ionization theory upon which the former value is based. It appears that a combination of two physical conditions may exist in the corona which may produce this discrepancy:

1. An average time between ionizations or recombinations which is large compared with the time for equipartition of energy between the ions and the coronal electrons;
2. A mass motion of ions through a region having a high electron temperature gradient in a period of time equal to or less than the above mentioned time between ionizations.

First consider the existence of each of these effects in the lower corona and their influence on a theory with ionization equilibrium assumed. An estimate of the time between ionizations can be obtained using values of the ionization cross-section for iron ions given and discussed by Firor and Zirin.[7] These cross-sections are found to increase only slightly with temperature (from 0.21×10^{-12} cm³/sec at 1.0×10^{6} °K to 6.4×10^{-12} cm³/sec at 2.6×10^{6} °K for Fe XII). The change with stage of ionization at a given temperature is still less sensitive, decreasing, for example, from 5.3×10^{-12} cm³/sec for Fe X to 2.0×10^{-12} cm³/sec for Fe XIV at 2.0×10^{6} °K. A measure of the time between successive ionizations is then given by

$$\tau_{\text{ion}}=N_i/N_iN_eS(T)=1/N_eS(T)$$

where N_i is the ion density, N_e is the electron density and $S(T)$ is the ionization rate in cm³/sec. Taking a value for N_e of 3.5×10^{8}/cm³ at a height of 1.02 $R\odot$ (where $R\odot$ is the solar radius) and a $S(T)$ of 3.0×10^{-12} cm³/sec we find that a time of the order of 1000 sec is required to ionize (without recombinations) an assemblage of ions from stage i to stage $i+1$. On the other hand, the time required to establish equipartition of energy between electrons and heavy ions in the coronal plasma is considerably shorter. This characteristic time can be computed from an equation given by Spitzer[8] and is of the order of 20 sec for Fe XIII-electron equilibrium at an electron density of 3.5×10^{8}/cm³ and temperature of 2.0×10^{6} °K. The time for equipartition is again inversely proportional to the electron density, which leads to the relation

$$\tau_{\text{ion}}/\tau_{\text{equi}}\approx\text{constant}\gg1$$

for a given ion anywhere in the solar corona.

In considering the second condition stated above we must restrict ourselves to regions of low magnetic field, thereby permitting an expansion of the corona which results in the corpuscular stream deduced by Biermann[9] from his work on comet tails and was more recently discussed by Parker.[10] Using the continuity equation we can compute the average solar wind velocity at any height in the corona once its velocity and particle density at another height have been defined. Using as initial parameters[11] for $r=1.43R\odot$, a proton density of $1.8\times10_1$/cm⁷ and a solar wind velocity of 66 km/sec, we find at $1.02R\odot$, where the density is approximately 3.5×10^{8}/cm³, a velocity of about 7 km/sec. From these figures we may conclude that an ion may travel radially outward at a distance of 10^{3} to 10^{4} km in the lower corona before undergoing further ionization.

The result of fulfilling both of the above conditions is that an ion, in traveling along a positive electron temperature gradient or in crossing a

steep gradient representing a large change in electron temperature, very quickly attains an average thermal velocity characteristic of the higher temperature (condition 1) but carries with it, for a much longer time and radial distance, a history of any previous ionization equilibrium (condition 2). Under such circumstances no theory which postulates equilibrium between ionization and recombinations can furnish a valid value for the electron temperature of the region even if the predominent ionization mechanism is electron impact, and recombination is by radiative capture. A kinetic temperature derived from line broadening may still be valid, however. Because of the large increase in temperature over a distance of 10,000 km or less in the lower solar corona it appears that the situation described herein may well occur throughout the lower corona where most of the coronal radiation is observed.

At any height in the corona, we then must replace the equilibrium equation

$$N_i N_e S_i^{i+1}(T) = N_{i+1} N_e Q_{i+1}^{i}(T),$$

where $Q(T)$ is the recombination rate in cm^3/sec^1, with the more general equation

$$\nabla \cdot \mathbf{F}_i \gtrless 0$$

where $\mathbf{F}_i$ is the flux of ions in stage i of each element.

For an optically thin corona of low density in which photoionizations and three-body recombinations are assumed to be negligible we obtain

$$\nabla \cdot \mathbf{F}_i = N_{i-1} N_e S_{i-1}^{i}(T) - N_i N_e [Q_i^{i-1}(T) + S_i^{i+1}(T)] + N_{i+1} N_e Q_{i+1}^{i}(T),$$

for each stage of ionization i of each element present, together with

$$N(r) = \sum_i N_i(r),$$

where N is the number density in each stage of ionization. The summation is made over all stages of ionization of the particular element being considered.

For a spherically symmetric corona expanding with velocity of magnitude $v(r)$ we have, for the total flux of a particular ion crossing a shell of radius r,

$$F_i = 4\pi r^2 N_i(r) v(r).$$

By requiring a steady state outward flow (i.e., all ions moving outward with the same average velocity), we obtain

$$F_0 = \sum_i F_i \approx 4\pi r^2 N_e(r) v(r) = \text{constant}$$

In this case the summation includes all elements present.

Using the notation P_i for the relative population in each stage of ionization,

$$P_i = N_i(r)/N(r),$$

we obtain

$$(d/dr)P_i = \frac{N_e^2}{F_0/4\pi r^2}\{P_{i-1} S_{i-1}^{i}(T) - P_i[Q_i^{i-1}(T) + S_i^{i+1}(T)] + P_{i+1} Q_{i+1}^{i}(T)\}$$

for each element and each stage of ionization.

It is possible to discuss the general behavior of the system by considering the expansion of an assemblage of ions from a region of low electron temperature 10^5 °K to a hotter region of 10^6 °K, the change in electron temperature taking place over a distance of only a few thousand kilometers. Examination of the equation for the lowest stage of ionization present at 10^5 °K indicates that the relative population will decrease, in the absence of recombinations, by a factor of $1/e$ in a distance of

$$\frac{F_0/4\pi r^2}{N_e^2 S(T)}$$

which is of the order of 10^4 km. Including recombinations will increase this distance. Higher stages of ionization will become more populated with increasing height above the photosphere, reaching an ionization equilibrium characteristic of the higher temperature only in a distance of a few tens of thousands of kilometers. It is precisely in this height range that most of the coronal radiation is observed.

Comparison with Observations

A small amount of observational evidence is available which appears to support the scale

computed. Beginning with the lower stages of ionization of iron observed in the corona, it has been found for the undisturbed corona that:

1. The maximum in Fe XI emission occurs below 10,000 km above the photosphere (February 25, 1952 eclipse).[12]
2. The maximum in Fe XIV occurs (statistically) at about 18,000 km.[13]
3. The intensity of the Fe XV line at 7059A decreases much more slowly with distance from the limb than do the lines of Fe X (6374A) and Fe XI (7891A) (February 15, 1961 eclipse).[14] Based on the observation by OSO–1 that the intensity of emission from Fe XVI requires emission from 0.1 to 0.5 of the solar corona we might hypothesize, in accordance with the model being presented here, that the maximum in emission of Fe XVI occurs at about 40,000 km. Because of the rapidly decreasing electron density with height in the lower corona the maximum in intensity of an ion may be very broad and need not be coincident with the maximum in relative population.

It is also of interest to reverse the problem by using the experimental data to estimate the radial velocity that would be required to achieve the observed separations of maximum emission assuming again that excitations are produced primarily by electron collisions. Using an average time between ionizations of 1000 sec for stages Fe XI through Fe XIV and 2000 sec for FeXV and Fe XVI we obtain:

For the radial velocity of Fe XI through Fe XIV,

$$\bar{v}\approx 10{,}000/3000 = 3.3 \text{ km/sec};$$

For the radial velocity of Fe XV and Fe XVI,

$$\bar{v}\approx 20{,}000/4000 = 5 \text{ km/sec}$$

These values are comparable to the solar wind velocity of the lower corona computed on page 7. By comparison, the average thermal velocity for Fe ions at 2.0×10^6 °K is 16 km/sec. Since only limb observations of the corona are possible at visible wavelengths, little or no shift in the emission lines, due to their outward velocity, should be observable.

SUMMARY

Intensity variations of three spectral lines observed in the solar EUV spectrum have been presented together with a discussion of the slowly varying component of solar activity which accompanied these changes. The Fe XV and Fe XVI lines are found to be prominent in the solar spectrum even in the absence of solar activity—a fact not expected on the basis of observations of forbidden lines of iron in the visible spectrum. It is found that a general correlation with solar activity exists, with the suggestion that an increase in these coronal radiations occurs subsequent to appearance of a center of activity as observed at 2800 Mc and may linger in the presence of diminishing Ca+ plages after the cessation of solar activity in an active region. The enhancement at 284A and 335A is nearly the same which implies little or no temperature increase in active regions having no yellow line (Ca XV) emission. These observations are consistent with a model for active regions which postulates both an increase in electron density and an increase in the vertical range over which radiation from Fe XV and Fe XVI is emitted. The discrepancy observed between the electron temperature obtained from ion ratios and that inferred from electron density profiles for the quiet corona has led to the suggestion that the former method may yield a fictitiously low value if there is a streaming outward of ions from the base of the corona. An approximate calculation shows that a radial velocity outward (small compared with the average thermal velocity of the ions) is sufficient to explain the observed increase in the maximum height of emission of the Fe ion sequence with increasing stages of ionization.

REFERENCES

1. Hall, L. A., Damon, K. R., and Hinteregger, H. E., "Solar Extreme Ultraviolet Proton Flux Measurements in the Upper Atmosphere of August 1961," in Space Research III, ed. by W. Priester, Amsterdam; North-Holland Publishing Co., pp. 745–759, 1963.
2. deJager, C., "The Development of a Solar Center of Activity," in "Vistas in Astronomy," v. 4 (A. Beer, ed.), 143–183, New York: Pergamon Press, 1961.

3. WOOLLEY, R. v. d. R., and ALLEN, C. W., "The Coronal Emission Spectrum," *Roy. Astron. Soc. Mon. Not.*, **108**(3): 292–305, 1948.
4. ELWERT, G., "The Continuous Emission Spectrum of the Solar Corona in the Far U.V. and the Adjacent X-Radiation" (in German), *Z. Naturforsch.*, **7a**, 202–204, February 1952.
5. ———, "Ionization and Recombination Processes in a Plasma, and the Ionization Formula for the Solar Corona" (in German), *Z. Naturforsch.*, **7a**, 432–439, June 1952.
6. VARSAVSKY, C. M., "Some Atomic Parameters for Ultraviolet Lines," *Astrophys. J. Suppl. Ser.*, **6**(53): 75–108, March 1961.
7. FIROR, J., and ZIRIN, H., "Observations of Five Ionization Stages of Iron in the Solar Corona," *Astrophys. J.*, **135**(1): 122–137, January 1962.
8. SPITZER, L., Jr., "Physics of Fully Ionized Gases," New York: Interscience Publishers, 1956.
9. BIERMANN, L., "Remarks on the Law of Rotation in Terrestrial and Stellar Zones of Instability," *Z. Astrophys.*, **28**(3): 304–309, 1951.
10. PARKER, E. N., "The Hydrodynamic Theory of Solar Corpuscular Radiation and Stellar Winds," *Astrophys. J.*, **132**(3): 821–826, November 1960.
11. PARKER, E. N., "The Solar Wind," in "Space Astrophysics" (W. Liller, ed.), 157–170, New York: McGraw-Hill Book Co., 1961.
12. ATHAY, R. G., and ROBERTS, W. O., "Coronal Line Intensities at the Khartoum Eclipse," *Astrophys. J.*, **121**(1): 231–240, January 1955.
13. PETRI, W., Forschungsber. Obs. Wendelstein No. 2, 1952.
14. WLERICK, G., and FEHRENBACH, C., "Premiers Resultats Concernant les Spectres Infrarouges de la Couronne Obtenus pendant l'eclipse du 15 Fevrier 1961," *Proc. I.A.U.* Symposium 16, 1963.

THE SOLAR SPECTRUM FROM 50A TO 400A

W. M. NEUPERT, W. E. BEHRING AND J. C. LINDSAY

Goddard Space Flight Center

A scanning monochromator, mounted as a pointed experiment on OSO–1, has been used for observations of solar extreme ultraviolet radiation from 50A to 400A. The period of observation was sufficient to observe a slowly varying component having a period of approximately 27 days and correlating with solar activity. The enhancement of radiation during periods of activity was observed to vary from line to line in the spectrum, depending upon the origin of the line in the solar atmosphere. Data showed an increase (33%) in the He II Lyman-alpha (304 A) flux, integrated over the entire solar disk during a period when the Zurich Provisional Relative Sunspot Number increased from 0 (March 11, 1962) to a maximum of 94 (March 22, 1962). Enhancements of approximately a factor of 4 were observed for the 284A (Fe XV) and the 335A (Fe XVI) lines.

INTRODUCTION

A grazing incidence spectrometer has been flown as a pointed experiment on OSO–1 (1962 ζ1) to monitor the extreme ultraviolet spectrum of the sun. Over 6,000 spectra were obtained at the rate of about 100 per day from March through May 1962, corresponding to nearly 3 solar revolutions. Intermittent operation has subsequently provided spectra more than 1 year. This paper presents, in a condensed form, the variations observed in three of the more reliable identified lines of the spectrum: 284A (Fe XV), 304A (He II) and 335A (Fe XVI).

The schedule of reduction of telemetry records was designed to provide a coarse look (several orbits per day) throughout the 3-month period of continuous observation. This approach is not suited for the detailed study of a particular event for which one must have continuous data over a period of hours. Therefore, transient phenomena on the sun are only now being studied as data reductions become more complete and the results will be presented in a future paper.

DESCRIPTION OF SPECTROMETER

During operation the spectrometer was pointed at the center of the solar disk, within approximately 2 minutes of arc. At this orientation, radiation from the entire solar disk and inner corona passed directly through the entrance slit and struck a concave grating mounted in grazing incidence, the angle of incidence being 88°. The grating, an original ruled in a special glass by the Nobel Institute in Stockholm, had 576 lines per millimeter on a blank with 1 m radius of curvature. The exit slit and detector were mounted on a carriage which was driven on a circular rail so that the slit scanned along the Rowland Circle, where the spectrum was focused, from 10A–400A. The plane of the exit slit was approximately perpendicular to the diffracted ray at all positions along the track, thereby keeping the spectral passband nearly constant for all angles of diffraction. The 50μ entrance and exit slits provided a spectral passband of 1.7A and permitted resolution of lines 0.85A apart. The detector was a windowless photomultiplier developed by the Bendix Corp. specifically for use in this spectrometer. A tungsten photocathode was chosen to minimize response to wavelengths above 1500A, and to reduce changes in sensitivity due to variations in the emission properties of the cathode.

CALIBRATION OF THE SPECTROMETER

The instrument was calibrated by: (1) exposing the entire spectrometer to a beam of mono-

chromatic radiation of known intensity; or (2) by evaluating the essential components of the spectrometer (grating, detector, etc.), and then computing the sensitivity of the instrument. The first method was used at 44A using a proportional counter for determining the source intensity. The second method was applied at longer wavelengths (80A–400A).

In addition, a comparison was made of the solar fluxes obtained against the fluxes measured by Hall, Damon, and Hinteregger[1] with a calibrated rocket instrument. The comparison could only be made in the region of overlap, 250A to 400A, and would be meaningful only if the solar radiation was the same. The 2800 Mc mean daily flux recorded by the National Research Council, Ottawa, Canada, was used as an independent estimate of solar flux to choose the satellite data for the comparison.

A best fit was made between these methods of obtaining a calibration and yielded values of 8.0×10^5 photons/cm^2-count at 335A and 4.2×10^5 photon/cm^2-count at 284A.

PRESENTATION OF DATA

Figure 1 presents a typical scan obtained over the region, 170A to 400A. Several factors reduce the usable spectral range of the spectrometer from its nominal range of 10A to 400A. At wavelengths below 100A the decreasing sensitivity of the spectrometer combined with an increase in scattered light upon approaching the central image made it impossible to distinguish a reliable spectrum on a single scan. However, a combination of several scans, using cross-correlation techniques, has shown the apparent existence of a line emission spectrum in the region from 50A to 100A. At wavelengths above 342A the second order images of intense spectral lines (observed in the first order above 171A) obscure a considerable amount of interesting data. In figure 1 this effect is indicated by cross-hatching those first order lines which have strong second order counterparts in the recorded spectrum. These second order images were less prominent in earlier rocket results, thus there is considerable dissimilarity between the rocket and satellite spectra above 342A.[2] For these reasons the obviously usable range of the instrument can be considered to be from 170A to 340A–a range which can be extended, with special precautions, to cover the interval from 50A to 400A.

The brightest emission line in the region from 170A to 340A is the Lyman-alpha line of ionized helium at 304A. In addition, numerous other emission lines appear with combined flux comparable to, or somewhat greater than, that of the helium line. Resonance liens of heavy ions (Mg through Fe) are expected in this region, leading to attempts[2,3] to identify the more prominent features of the spectrum in terms of such lines. The difficulty of positive identification can be appreciated more if, for example, we realize that the resonance lines of all stages of ionization of

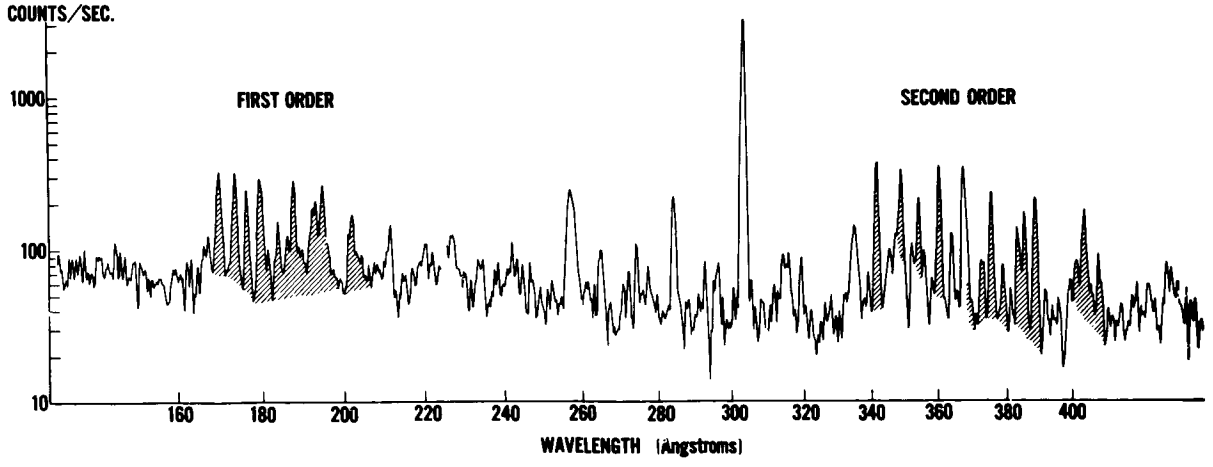


FIGURE 1.—Solar spectrum from 170A to 400A with second order line images indicated.

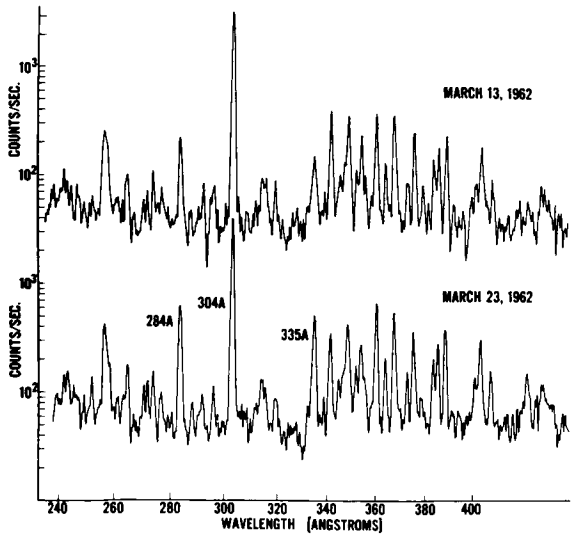


FIGURE 2.—Comparison of two spectral scans taken under different solar conditions.

iron from Fe X through Fe XVI, with the exception of Fe XV, are expected in the small wavelength interval from 335A to 390A.[2] The overlapping of these resonance multiplets combined with second order contamination of shorter wavelengths makes positive identification a formidable task. The resonance lines of Fe X through Fe XIV have not been identified in the extreme ultraviolet (EUV) spectrum although their intensity as predicted by Ivan-Kholodny and Nikolsky[4] should permit their observation. Tousey,[5] working with a spectrum having higher resolution, reports that no Fe XIV multiplet can be found in his record. Because of these difficulties, discussion will center around three of the more intense and more reliably identified lines of the spectrum: the Fe XV ($3s^2\ ^1S_0-3s\ 3p\ ^1P^0_1$) line at 284A; the Fe XVI ($3s\ ^2S_{1/2}-3p\ ^2P^0_{3/2}$) line at 335A; and, in addition, the He II ($1s\ ^2S_{1/2}-2p\ ^2P^0_{3/2,1/2}$) line at 304A. The location of these lines is indicated in figure 2.

TIME VARIATIONS OF THE EUV SPECTRUM

The months of March and April 1962 were auspicious for a study of the solar EUV spectra in that observations could be made on both a quiescent and a disturbed solar atmosphere. During the second week in March the sun was especially quiet—the sunspot number being 0 on March 11. As the month progressed the solar rotation carried several centers of activity across the visible hemisphere of the sun. Definite enhancements in the solar spectrum were associated with these centers of activity.

Figure 2 presents two scans of the EUV spectrum which were obtained approximately 10 days apart. During the first of these observations (March 13) only one small region of activity was present on the solar disk. In spite of this low level of activity it is observed that the Fe XV and Fe XVI lines persist as two of the more prominent features of the spectrum. The second spectrum (March 23) was obtained while several large and well developed centers of activity were present on the disk. Comparing these two spectra we observe that the emission lines have increased in intensity but not all by the same amount. The Fe XV and XVI lines, already prominent even in the absence of solar activity, have increased in intensity appreciably more than any other line observed with certainty in this spectral range. The He II line has also increased, but by a lesser amount.

Time variations of these three lines as observed for the first 1,066 orbits of operation are given in figure 3, in which each datum point represents the average of 1 orbit's observations (6 to 8 scans).

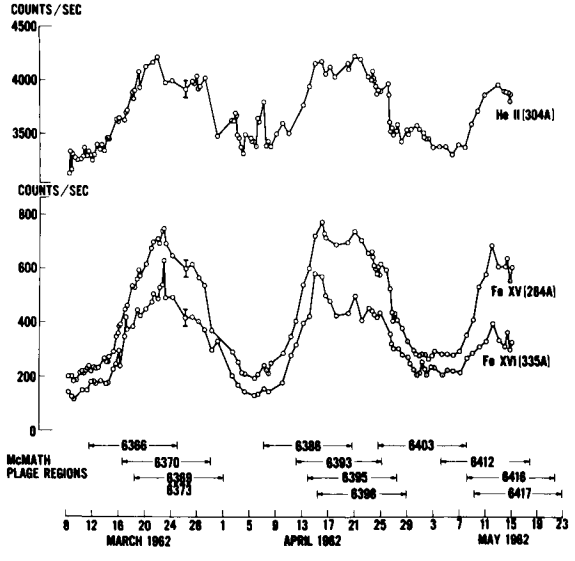


FIGURE 3.—OSO-1 data for three spectral lines compared with the appearance and disappearance of major plage areas.

Error bars indicating one standard deviation in the data are given for each wavelength. Also indicated are the times of appearance and disappearance of the major plage areas as cataloged by the McMath–Hulbert Observatory. The first increase in the counting rates above the "quiet sun values is associated with the appearance of McMath Plage No. 6366 on the east limb on March 11. The slope of the He II curve is observed not to change appreciably on March 17 as several more prominent plages appear on the limb, although their appearance is obvious in the radio data. The counting rate for the He II (304A) line increases gradually to a maximum on March 23, after which it drops rather abruptly to an apparent plateau. This sudden drop cannot be unambiguously attributed to the disappearance of plage region 6366 on the west limb, for it follows quite directly after an unusual flare of importance 3 observed in plage region 6370 on March 22, for which data are not given in figure 3. A gradual decline in readings is observed from March 29 to April 7. A similar enhancement, followed by a return to near preplage rates, occurred in April. Since the spectrometer did not carry an internal standard (e.g., radioactive source) it is impossible to state positively that its counting characteristics did not change over the period of observations. The consistent behavior of the data, however, as demonstrated in figure 3, leads to the conclusion that no significant change in sensitivity occurred throughout the period of observation. (Subsequent analysis of real time data after 1 year in orbit supports this conclusion, even for the longer time interval.)

Also shown in figure 3 are the variations with solar activity of the coronal Fe XV (284A) and Fe XVI (335A) lines, the relative increases being considerably larger than for the He II (304A) line. A definite, although small, enhancement associated with the appearance of plage No. 6366 is observed in both the Fe XVI and Fe XV lines. The possibility exists that the coronal emissions appear slightly earlier as might be expected since the coronal emission must take place at greater heights than the calcium radiation. The most obvious increase in intensity of these lines is associated with the appearance, in succession, of plage areas numbered 6370, 6369, and 6373 on March 17

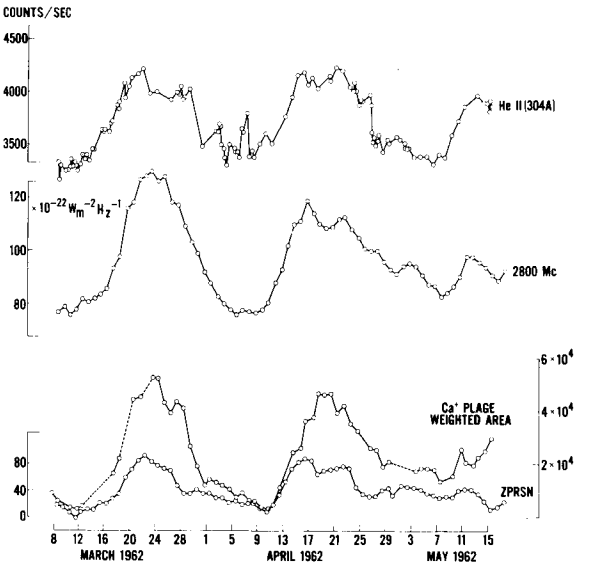


FIGURE 4.—Relationship of observed counting rates for the He II Lyman-alpha line to ground-base measurements of solar activity.

through March 19. The counting rates then increased gradually throughout this period indicating that the enhanced emission was not from a point source but rather from an extended volume having as its base an area at least as large as the underlying calcium plage. The disappearance of area 6366 on March 25 results in a slight decrease in counting rates, followed by a general decline from March 29 through April 3 as the other plage areas are carried off the visible hemisphere of the sun by the solar rotation.

Although the horizontal scale in figure 3 is inadequate for the display of transient phenomena, at least one interesting event, associated with a flare of importance 3, on March 22, can be discerned. This flare was unusual in that the ionospheric effects produced by it were more nearly characteristic of a small (importance 1) flare. An enhancement was observed at 304A (not plotted) simultaneously with the visible flare, but not at 335A or 284A. However, at the latter wavelengths a significant enhancement was noted one or two orbits after the visible maximum. The peak point at 335A lies at eight standard deviations from the mean of the day's observations before the flare and represents a nearly certain change in the EUV flux. Further analysis of this

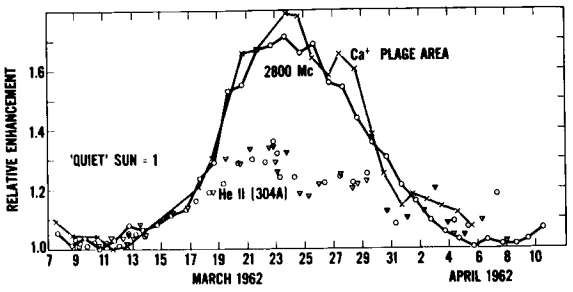


FIGURE 5.—Comparison of 304A data, normalized to a "quiet sun," with calcium plage data and 2800-mc radio observations.

event covering the major lines of the spectrum will be presented in a future paper.

The relationship of these observed counting rates to several ground-based measurements of solar activity is presented in figures 4 and 6. In figure 4, the He II radiation is compared with daily values of the solar flux at 2800 Mc, and with the Zurich Provisional Relative Sunspot Number (ZPRSN). Also shown is an estimate of the calcium plage area, each area being weighted by the estimated intensity of the area on a scale from 1 to 5. Values for this computation were supplied by the McMath–Hulbert Observatory.

A more detailed comparison is made in figure 5 in which the radio and He II fluxes have been adjusted by dividing each observation by the rate observed for the "quiet sun." In addition, the Ca^+ plage intensity has been adjusted to coincide with the 2800 Mc enhancement on March 21. This plot shows more clearly the close relationship

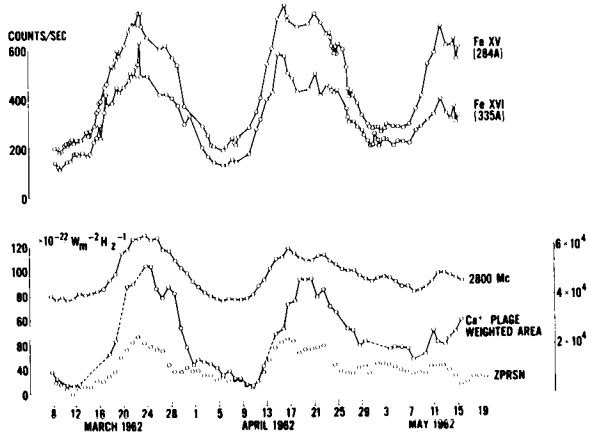


FIGURE 6.—Relationship of observed counting rates for the 284 Aand 335A lines to ground-based observations.

which has been known to exist for some time between the 2800-Mc radiation and the plage areas. The divergences of the He II (304A) line from both of these curves is also shown. In figure 6 the daily values of solar flux at 2800 Mc and the Zurich Provisional Relative Sunspot Number are compared with radiation due to the coronal lines of Fe XV (284A) and Fe XVI (335A). The estimated calcium plage intensity is also given.

DISCUSSION

The spectral lines chosen for presentation here were selected because they are reliably identified with particular ions, not because they convey better than any other line in the spectrum the plage-associated changes in solar flux. In terms of fractional changes in intensity, these three lines represent the extremes which have thus far been observed in the spectral region from 171A to 400A; only a few faint lines have smaller nonflare variations than the He II Lyman-slpha line, while no other lines have increases as great as those observed for 284A and 335A. A summary of the average increases in counting rates for the period from March 9 to March 23 1962, a period of increasing solar activity, is given in table 1. The increase, weighted by the intensity of each line, is computed for the range from 171A to 305A, using 60 reliably observed lines. The increase in the range from 305A to 400A can only be estimated because of the masking effect of second order images above 342A. The values given in table 1 are, of course, appropriate only for the particular interval of time for which they were computed.

TABLE 1.—*Increases in Solar EUV Spectrophotometer Counting Rates, 9 March 1962 to 23 March 1962*

Spectral Range (A)	Average Counting Rate Increase (Percent)
171–228	55.
229–300	80.
229–305	52.
305–400	50 (estimated).

The initial analysis of only three lines (He II 304A, Fe XV 284A, Fe XVI 335A) already indicates that the relative prominence of spectral lines may depend upon the age of the center of activity which is responsible for the increased radiation. As an example, the maximum emission in the Fe XV (figure 6) apparently occurs later in time than the maximum for the 2800 Mc radio flux or for the plage areas observed during March, April, and May. In addition to such a slowly changing effect, one may note that localized perturbations appear (March 7–9 and April 16–17) for which the relative increases are considerably different for the helium and the iron lines. It appears that in these instances we are observing phenomena localized at particular levels in the solar atmosphere.

CONCLUSIONS

Observations of the solar spectrum between 170A and 340A have been made over a time interval corresponding to approximately three solar rotations. The observations have shown:

1. That the He II (304A) emission is enhanced by a factor of about 33 percent during a period when the Zurich Provisional Relative Sunspot Number increased from 0 to a maximum of 94 and the 2800 Mc flux varied from approximately 76×10^{-22} to 125×10^{-22} w/m^2-Hz.

2. The Fe XV (284A) and Fe XVI (335A) coronal lines were enhanced during the same period by a factor of approximately 4.

3. The enhancement of He II (304A) and Fe XV (284A) and Fe XVI (335A) due to plage activity was larger than enhancements due to flares that occurred during the 3-month interval of the observations.

4. The variations in intensity of the He II (304A), Fe XV (284A) and Fe XVI (335A) represent the extremes observed. If 60 of the reliably observed lines between 171A and 342A are averaged, the enhancement is between 50 and 80 percent for the time interval March 9 to 23.

5. Although there appears to be a gross correlation between solar activity indices (such as 2800 Mc flux) and the He II, Fe XV and Fe XVI fluxes, there are indications that the relative prominence of the spectral lines may depend upon the age of the center of activity.

REFERENCES

1. Hall, L. A., Damon, K. R., and Hinteregger, H. E., "Solar Extreme Ultraviolet Photon Flux Measurements in the Upper Atmosphere of August 1961," Space Research III, ed. by W. Priester. Amsterdam: North-Holland Publishing Co., 1963, pp. 745–759.
2. Neupert, W. M., and Behring, W. E., "Solar Observations With a Soft X-Ray Spectrometer," *J. Quant. Spectr.* Radiat. Trans. **2,** 527–532, October–December 1962.
3. Zirin, H., Hall, L. A., and Hinteregger, H. E., "Analyses of the Solar Emission Spectrum From 1300A to 250A as Observed in August 1961," Space Research III, ed. by W. Priester. Amsterdam: North-Holland Publishing Co., 1963, pp. 760–771.
4. Ivanov-Kholodnyi, G. S., and Nikol'skii, G. M., "A Prediction of Solar Line Emission in the Extreme Ultraviolet" (in Russian), *Astron. Zh. USSR,* **38**(5): 828–843, 1961.
5. Tousey, R., "The Extreme Ultraviolet Spectrum of the Sun," *Space Sci. Rev.* **2**(1): 3–69, July 1963.

SOLAR X-RAYS: SLOW VARIATIONS AND TRANSIENT EVENTS

WILLIAM A. WHITE

Goddard Space Flight Center

Solar X-ray flux integrated over the interval 0.1–10A was measured from the OSO–1 satellite in early 1962 using a xenon-filled ion chamber with a thin beryllium window. A slowly varying component of X-ray flux was observed which correlates well with the slowly varying component of 2800 mc solar radiation. The component of the X-ray flux can be accounted for by postulating localized sources having the same horizontal extent as Ca plages with thicknesses proportional to their diameter; and having an electron temperature of about 2.8×10^6 °K and an electron density of about $5(10)^9$ electrons/cm^3. A further conclusion is that for these conditions the ratio of line emission to continuum emission is at least 10:1 and more probably 30:1.

In addition to a slowly varying component, transient events (X-ray flares) lasting from 10 minutes to a few hours were frequently observed. Correlations with $H-\alpha$ flares, with SID's, and with 2800 mc transients have been investigated; the results show that as an indicator of local solar activity, the OSO–1 X-ray experiment was more sensitive by a large factor than indicators based on ionospheric effects or than indicators based on observations of solar flux in visible or radio wavelengths. X-ray flares were frequently observed to be associated in groups possessing a characteristic pattern; the implications are discussed.

INTRODUCTION

With the launching of the OSO–1 (1962 ζ1) on March 7, 1962, it became possible for the first time to point instruments at the sun accurately and continuously for entire daylight portions of a satellite orbit; for the 550-km orbit of OSO–1 these observing time intervals were the order of 1 hr, separated by darkness intervals of about $\frac{2}{3}$ hr. The observing periods were long enough to disclose some interesting dynamic effects which which would be difficult to study otherwise.

SENSOR CHARACTERISTICS

OSO–1 provided coverage of the solar X-ray radiation near a wavelength of 10A by means of an ion chamber whose characteristics are given in table 1. The conversion efficiency as a function of wavelength is shown in figure 1.

The full scale sensitivity is, of course, dependent upon the shape of the input spectrum. It will be shown that for nonflare periods it is reasonable to assume a spectral shape consistent with a 2.8×10^6 °K plasma; for such a spectrum, wavelengths contributing to the output current lie in the interval 3–11A; the full-scale sensitivity at such times is 1.8×10^{-3} erg/cm^2-sec within this wavelength region. For comparison with earlier measurements[1] over bandwidths specified as 2–8A, the full-scale sensitivity over a band is 3.6×10^{-4} ergs/cm^2-sec.

TABLE 1.—*Sensor Characteristics*

Characteristic	Value
Window material	Beryllium.
Window thickness	0.005 inch.
Total window area (two chambers in parallel)	3.38 cm^2.
Absorbing gas	Xenon.
Gas pressure	780 mm.
Ion chamber depth at normal incidence	2.19 cm.
Ion Pairs/erg	2.8×10^{10}.

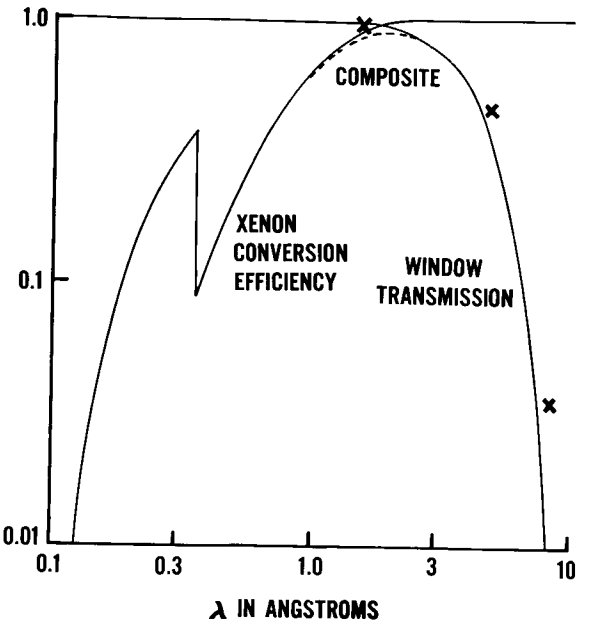


FIGURE 1.—OSO–1 ion chamber conversion efficiency as function of wavelength: The solid line represents the theoretical values and the crosses experimental data.

INPUT SPECTRUM ASSUMPTIONS

Both of the full-scale flux values mentioned above are computed for an input spectrum shape equivalent to a 2.8×10^6 °K plasma. The basis for this choice is threefold:

1. A direct measurement of spectral shape between 7 and 11A was made by Pounds, Willmore, et al.[2] from the satellite Ariel (1962 σ1) on April 27, 1962, at 2110 UT, 2 hours prior to a small visible flare. The spectrum obtained by this group fits a 2.8×10^6 °K plasma; integrating their absolute flux values between 7 and 11A gives a value for the integral of 1.2×10^{-3} ergs/cm^2-sec. At this time the OSO–1 X-ray ion-chamber experiment was off-scale, implying an integrated flux over the same wavelength interval (that is, 7–11A) of $\geq1.7\times10^{-3}$ ergs/cm^2-sec^{-1}, for the same spectral distribution. The agreement between OSO–1 and Ariel for the solar X-ray flux of April 27, 1962 is thus probably within a factor of 2.

2. An assumed temperature appreciably less than 2.8×10^6 °K would require too high an electron density to produce the X-ray fluxes measured by OSO–1. This statement is true either if the X-ray source is spread uniformly over the entire corona or localized in centers of activity. If we compute the continuum flux to be expected at these wavelengths from the entire solar corona (assumed isothermal at a somewhat lower temperature, say 2.4×10^6 °K) we find that to produce the lowest flux value measured by OSO–1, the integral of the square of the electron density taken over the entire corona is 55.3×10^{49}. This value should be compared with the value of $4.6(10)^{49}$ obtained by Shklovskii[3] using the coronal model of Allen,[4] or with the value of $3(10)^{49}$ obtained by Elwert.[5,6] Thus the theoretical flux in the continuum for an isothermal corona with Allen's electron density profile at 2.4×10^6 °K falls short of the lowest value measured on OSO–1 by a factor of order 15. In fact, we must place the entire corona at a temperature in excess of 3.5×10^6 °K to meet the lowest OSO–1 flux using such an all-continuum model. If the contribution from line emissions were allowed to exceed the flux from continuum emission by a factor of 15, the corona in its entirety would have to be at a temperature of about 2.4×10^6 °K; but this, remember, is for the lowest flux measured by OSO–1. For more than 50 percent of the time the OSO–1 flux exceeded this lowest value by at least a factor of 10.

3. An assumed temperature appreciably greater than 2.8×10^6 °K would be inconsistent with concurrent OSO–1 observations of the Fe XV to Fe XVI ratio made by Neupert.*

SPATIAL DISTRIBUTION OF X-RAY SOURCES

From the fact that most of the time the measured flux from OSO–1 was much larger than the lowest value measured (which is already uncomfortably high for an isothermal corona with uniform density profile), we can only conclude that the previously observed spatial localization of sources of X-rays of somewhat longer wavelength must also exist for wavelengths less than 10A, and inquire as to the conditions of temperature and electron density likely to be found in such local densifications. The observations of Billings[8] show that occasionally temperatures as high as 4.2×10^6 °K and electron densities as high

*See pages 535–539.

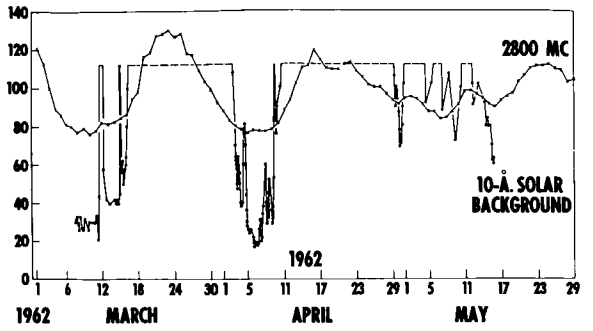


FIGURE 2.—Slowly-varying components of 2800 mc solar flux and of $\lambda < 11$A solar X-rays.

as 2×10^{10}/cm^3 are found; more usual values[9] run $T_e < 3.5\times10^6$ °K and $n_e < 7\times10^9$/cm^3.

SLOWLY VARYING COMPONENT

A comparison of the slowly varying part of the 10A X-rays with 2800 Mc radiation confirms that the localized sources of solar X-rays are in some way associated with centers of activity such as Ca plages and/or sunspot groups. Figure 2 shows the time history of both fluxes for about 2.5 solar rotations in the early life of OSO–1. It can be seen that the smoothed X-ray flux correlates fairly well with the excess 2800 Mc flux above a background of about 75 flux units appropriate for the "quiet" sun at that phase of the solar cycle.[8]

The lowest X-ray flux (which may be considered an upper bound on the X-ray flux from the quiet sun) measured by OSO–1 (on April 6, 1962) for $\lambda < 8$A was 3.6×10^{-5} erg/cm^2-sec; for $\lambda < 11$A, 1.8×10^{-4} ergs/cm^2-sec. This flux occurred at a time when only three plages of area ≥ 1000 millionths of a solar hemisphere were visible on the disk. The nearest plage behind the west limb had set 3 days previously, and the nearest behind the east limb was not to rise until 1.5 days later. If we now assume that the X-ray emission is coming from three "pillbox" volumes, one associated with each of the three plages, each having its base area equal to the plage area and furthermore having a height equal to one-half the diameter of its base, we have the situation shown in table 2. Using the total source volume obtained from table 2, table 3 gives the value of $\int n_e^2\, dv$ required to fit the observed X-ray flux of April 6, 1962 for several assumed temperatures, and shows the resulting values of n_e for various assumptions

TABLE 3.—*Electron Density for Plages of April 6, 1962*

T_e (10^6 °K)	$\int n_e^2\, dv$ (if a continuum)	n_e(10^{10} electrons/cm^3) Ratio of line emission to continuum			
		1:1	3:1	10:1	30:1
2.4	55.3×10^{49}	2.68	1.90	1.14	0.68
3.0	11.9×10^{49}	1.25	0.88	0.53	0.32
3.5	4.9×10^{49}	0.80	0.56	0.34	0.20

TABLE 2.—*Plages for April 6, 1962*

Observed calcium plage		Assumed X-ray source		
McMath plage No.	Area (millionths of solar hemisphere)	Area (cm^2)	Height (cm)	Volume (cm^3)
6377	1600	4.90×10^{19}	3.95×10^9	1.94×10^{29}
6378	1000	3.06×10^{19}	3.12×10^9	0.954×10^{29}
6379	1000	3.06×10^{19}	3.12×10^9	0.954×10^{29}
Total volume				3.85×10^{29}

regarding the relative contribution of line emission.[6] The volume above plage 6379 was observed by Billings* on April 9, 1962 to have a faint continuum enhancement from which he estimated $n_e = 0.5 \times 10^{10}$. If this was indeed the correct electron density, and if (as is most likely from such a weak source[9]) the electron temperature was not greater than 3.0×10^6 °K, table 3 indicates the ratio of line emission to continuum emission was at least 10:1. If T_e was no greater than the 2.8×10^6 °K, the value obtained from the Ariel spectrum of April 27, the ratio of line emission to continuum emission must have been around 30:1.

X-RAY FLARES

Up to now we have discussed only the slowly varying component of the X-ray emission: in addition to these quasi-steady-state conditions, transient events (X-ray flares) lasting usually from 10 minutes to a couple of hours were frequently observed. Such an event is shown in figure 3, and should be compared with the quiet period of similar duration shown in figure 4. The particular event of figure 3 contains a total energy below 11A of 2×10^{27} ergs.

Correlation with H-α Flares

During the 9-day interval between launch and March 16, 1962 (at which time the rising of plage 6370 on the cast limb supplied enough X-ray emission to carry the experiment off-scale), approximately 60 X-ray flare events lasting from 10 minutes to 1 hour were seen, and 4 events were seen to last about 5 hours.

During this same interval (1620 UT March 7, 1962, to 1620 March 16, 1962) some 33 H-α

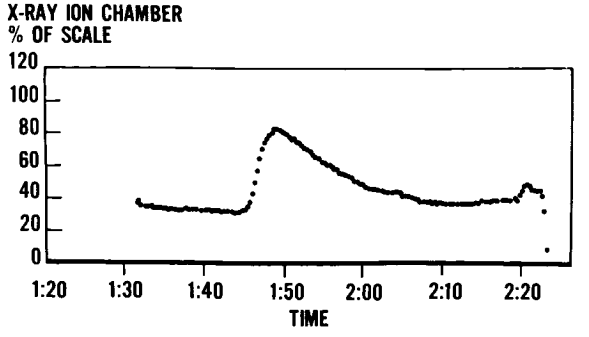


FIGURE 3.—Typical small X-ray flare.

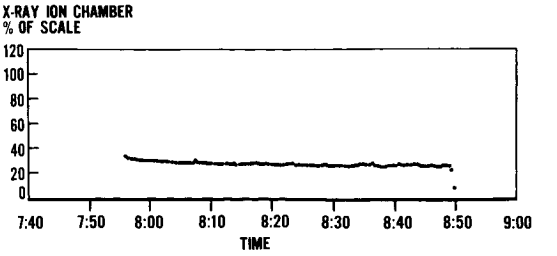


FIGURE 4.—Typical "quiet" period.

flares were reported by ground based observatories. Of these H-α flares, 6 would have been unobservable from OSO–1 for various reasons (satellite night, failure to command data storage readout, etc.). Of the remaining 27 H-α flares, 3 occurred while the X-ray experiment was still off-scale because of a previous large event. This leaves 24 H-α flares which can be tested for correlation with the X-ray flares. Of this group of 24, it appears that 11 correlate well, 3 definitely have no counterpart in X-rays, and the remaining 10 are doubtful because of insufficient data or an excessive time difference (greater than 10 minutes). Conversely, there are six full-scale or greater X-ray events for which no H-α flare was reported even though observations were presumably being made at the time. Certainly more observations will be required before a definite statement can be made regarding a correlation or lack thereof between H-α flares and X-ray flares.

Correlation with SID's

In looking for correlations with Sudden Ionospheric Disturbances (SID), all X-ray events exceeding the full-scale saturation level were barely detectable (if observing conditions permitted) in Sudden Phase Anomaly (SPA) data for VLF transmissions via the D layer. Only the large event of March 13 was seen in ionospheric indices other than SPA's.

Correlation with 2800 mc Transients

Correlation with transients in the 2800 mc solar flux is good; but again, full-scale X-ray events are represented by extremely small events (1–2 flux units) in the 2800 mc data.

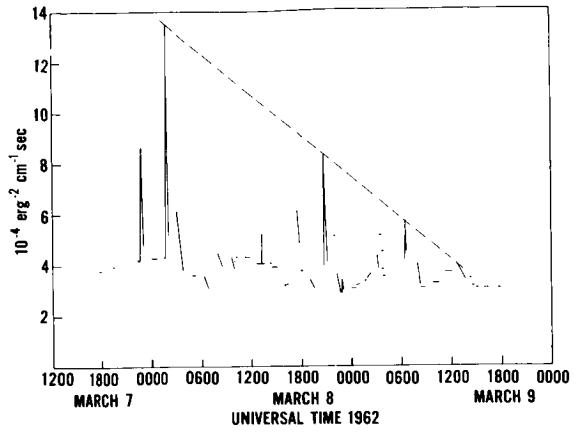


FIGURE 5.—Solar X-rays for $\lambda < 11$A March 7–9, 1962 showing grouping of X-ray flares.

Grouping of X-ray Flares

Several apparent associations of certain X-ray flares into groups displaying a definite pattern were observed; figure 5 shows such a grouping. Similar groupings are present in the data for the first week in April: in fact, the one particular March group shown in figure 5 has an exact April counterpart 27.1 days later with identical time separations between events and with identical peak excursions above mean background level. The envelope joining the peaks of the flares within a group is found: (1) to be a straight line, and (2) to have the same slope (with either positive or negative sign) from group to group.

These characteristics of the flare groups indicate time-rate-of-change of X-ray source strength. Source strength is a function of electron density, of temperature, and of volume: furthermore, it is difficult to see why any time variation of either density or temperature would be of such a particular nonlinear nature as to constrain the source strength to vary linearly with time. We are left with the concept of a *volume* which is either growing or diminishing at a constant rate, and which on occasion serves as a reservoir of high-temperature electrons and ions interacting to produce the X-ray flares.

SUMMARY

A slowly varying component has been found in the solar X-ray flux below 11A which correlates with the slowly varying component of the 2800 mc solar radiation. A model for these quasi-stable X-ray sources which fits the OSO–1 data postulates localized sources having the same horizontal extent as Ca plages with thicknesses proportional to their diameter, and having an electron temperature of about 2.8×10^6 °K and an electron density of about 5×10^9 electrons/cm^3. For these conditions it is also necessary that the ratio of line emission to continuum emission be at least 10:1 and more probably 30:1.

In addition to a slowly varying component, transient events (X-ray flares) lasting from 10 min to a few hours were frequently observed. Correlation with SID's and with 2800 mc transients has been obtained; correlation has been attempted with H-α flares with somewhat ambiguous results. X-ray flares were frequently observed to be associated in groups possessing a characteristic pattern; the concept of a source volume varying linearly with time is invoked to account for the linear envelope of a flare group.

ACKNOWLEDGMENTS

The author gratefully acknowledges the help of many of his associates at the Goddard Space Flight Center, and in particular, Alfred Stober and Robert Young who perfected the fabrication method for the ion chambers.

REFERENCES

1. KREPLIN, R. W., CHUBB, T. A., and FRIEDMAN, H., "X-Ray and Lyman-Alpha Emission From the Sun as Measured From the NRL SR–1 Satellite," *J. Geophys. Res.*, **67**(6): 2231–2254, June 1962.
2. POUNDS, K. A., WILLMORE, A. P., BOWEN, P. J., NORMAN, K., and SANFORD, P. W., *Proc. Roy. Soc.* (in press).
3. SHKLOVSKII, I. S., "The Ultraviolet Radiation and Soft X-Rays of the Sun," *Uspek. Fiz. Nauk*, **75**, 351–388, October 1961; Trans. in *Sov. Phys.-Uspek.*, **4**(5): 812–834, March–April 1962.
4. ALLEN, C. W., "The Spectrum of the Corona at the Eclipse of 1940 October 1," *Roy. Astron. Soc. Mon. Not.*, **106**(2): 137–150, 1946.
5. ELWERT, G., "The Soft X-Radiation From the Undisturbed Solar Corona," (in German), *Z. Naturforsch.*, **9a**, 637–653, July–August 1954.
6. ———, "Theory of X-Ray Emission of the Sun," *J. Geophys. Res.*, **66**(2): 391–401, February 1961.
7. BILLINGS, D. E., "Profile of the Yellow Coronal Line 5694," *Astrophys. J.*, **125**(3): 817–821, May 1957.
8. COVINGTON, A. E., and HARVEY, G. A., "The Visibility of the 10-cm Radio-Emissive Region and Its Application in Finding the 10-cm Quiet Sun," *Astrophys. J.*, **132**(2): 435–451, September 1960.
9. BILLINGS, D. E., Private communication.

IV. IONOSPHERE AND RADIO PHYSICS

AN INTERPRETATION OF A ROCKET MEASUREMENT OF ELECTRON DENSITY IN THE LOWER IONOSPHERE

A. C. AIKIN, JR., J. A. KANE

Goddard Space Flight Center

AND

J. TROIM

Norwegian Defence Research Establishment

A mid-latitude rocket measurement of electron density is presented for altitudes between 65 and 85 km. An interpretation of the profile indicates that the region between 70 and 85 km is ionized by the action of Lyman alpha radiation on nitric oxide rather than by 2–8A X-rays. Possible explanations are given for a minimum found in the electron density distribution in the vicinity of the mesopause.

INTRODUCTION

Current theories attribute the formation of the ionosphere below the mesopause to: (1) Ionization of the major atmospheric constituents by the highly variable 2–8A solar X-radiation; (2) Ionization of the minor constituent, nitric oxide, by Lyman alpha at 1215.6A; and (3) Cosmic rays, which extend the ionosphere below 70 km. The theoretical problem is further complicated by the possibility of a wide variety of electron loss processes including dissociative recombination of an electron-ion pair and the formation of negative ions.

A lack of quantitative data of the electron and ion density distribution and too few attempts to correlate observational results with the intensities of ionizing solar radiations, has made it difficult to check the validity of any theory. This paper will deal with the interpretation of a rocket measurement of electron density in the ionospheric D region. The profile was obtained at 1930 GMT on March 8, 1963 at Wallops Island, Va. for a solar zenith angle of 48°.

EXPERIMENTAL RESULTS

An electron density profile (fig. 1) was obtained using a Faraday rotation technique involving transmission of a 3 mc signal from the ground to the rocket. Each point is the average electron density in the height interval indicated by the vertical bar. The horizontal bar indicates the uncertainty in the determination of this average

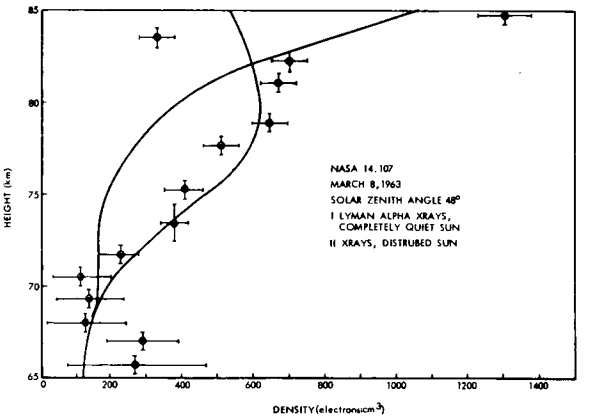


FIGURE 1.—Observed electron concentration in the D region and two theoretical profiles (solid curves).

value. This uncertainty is due to random echoes from above the D region distorting the Faraday pattern.

Below 75 km a knowledge of electron collision frequency is necessary to deduce electron densities from a 3 Mc Faraday rotation experi-

ment. The Faraday rotation experiment by itself allows the collision frequency to be determined at a single altitude. This follows from the fact that the sense of rotation of the plane of polarization is positive or negative for the collision frequency ν_m greater or less than the exploring frequency. By noting the altitude at which the reversal in the sense of the Faraday rotation is observed a value $\nu_m = 13.8 \times 10^6$/sec at an altitude of 61 ± 2 km is deduced. The value is consistent with measurements made at Fort Churchill by Kane.[1] Clearly discernible from the profile are: (1) The base of the E region, (2) The D region between 84 km and 70 km; and (3) A region below 70 km which has been designated as the C region. The minimum in electron density at 83.5 km coincides with the mesopause. Additional evidence for the existence of the minimum has been obtained on previous rocket flights from Wallops Island.

INTERPRETATION OF THE ELECTRON DENSITY PROFILE

In order to interpret the electron density profile, it is necessary to consider the equations governing the distribution of ions and electrons as outlined by Nicolet and Aikin.[2] Under conditions of equilibrium:

$$n(NO^+)n_e = \frac{n(NO)K_{iNO}Q_{L_\alpha}}{\alpha_{DNO}+\lambda\alpha_i}; \quad (1)$$

$$n(O_2^+)n_e = \frac{n(O_2)K_{iO_2}\epsilon Q_x + q_{O_2CR}}{\alpha_{DO_2}+\lambda\alpha_i}; \quad (2)$$

$$n(N_2^+)n_e = \frac{n(N_2)K_{iN_2}\epsilon Q_x + q_{N_2CR}}{\alpha_{DN_2}+\lambda\alpha_i}; \quad (3)$$

$$\lambda = \frac{n_-}{n_e} = \frac{1.5 \times 10^{-30}[n(O_2)]^2}{0.44}; \quad (4)$$

$$n(NO^+)+n(O_2^+)+n(N_2^+) = (1+\lambda)n_e. \quad (5)$$

The number of Lyman alpha or X-ray photons at a given height h is represented by Q_{L_α} and Q_x. These quantities should be measured simultaneously with the electron density. K_i is the ionization cross section and ϵ the efficiency of ionization. $q_{O_2\,CR}$ and $q_{N_2\,CR}$ represent the ionization rates due to cosmic rays and have been assigned the values of $10^{-17}\,n(O_2)$ and $10^{-17}\,(N_2)$, respectively. The quantities $n(O_2)$, $n(N_2)$ and $n(M)$ represent the number density cm^{-3} of molecular oxygen, nitrogen and the sum of all constituents. It will be assumed that the rates of dissociative recombination between positive ions and electrons of density, n_e, measured by Kasner, Rodgers and Biondi[3] are operative, namely $\alpha_{DN2} = 6 \times 10^{-7}$ cm^3/sec and $\alpha_{DO_2} = 4 \times 10^{-7}$ cm^3/sec. The effect of negative ions, n_-, is represented by λ which is the ratio between the rate of attachment of electrons to molecular oxygen and photodetachment. The coefficients involved were adopted from Chanin, Phelps and Biondi[4] and Burch, Smith and Branscomb.[5] Negative and positive ions recombine at a rate α_i, which is taken to be 10^{-7} cm^3/sec.

In order to determine the relative importance of 2–8A X-rays and Lyman alpha in the formation of the D region two cases are computed. One case (curve 1) involves the ionization by solar X-rays whose intensity is representative of a completely quiet sun together with Lyman alpha acting on nitric oxide. The value of the Lyman alpha flux as measured above the D region on a subsequent flight is 2.4×10^{11} photons/cm^2 sec with a possible error of 25 percent. Good agreement is found with the experimental electron density data if it is assumed that $\alpha_{DNO} = 4 \times 10^{-8}$ cm^3/sec and $n(NO) = 1.5 \times 10^{-10}\,n(M)$. However, these values are not unique, since it is only possible to determine the ratio $n(NO)/\alpha_{DNO}+\lambda\alpha_i$ from the data.

Table 1 illustrates the distribution of X-ray energy assumed. Pounds[7] indicates that the intensity values assumed for the completely quiet sun case are in agreement with measurements made during this portion of the sunspot cycle. The second case considers ionization solely by X-rays which are characteristic of a disturbed sun. Since only a solar flare, which was not observed, would give sufficient intensity to fit the experimental data the observed profile is strong evidence for the Lyman alpha hypothesis.

TABLE 1.—*X-ray Intensities for Various quiet and Disturbed Solar Conditions*

Wavelength (A)	X-ray intensity (erg/cm² sec)	
	Quiet sun	Disturbed sun
2	10^{-8}	10^{-6}
4	10^{-7}	10^{-5}
6	10^{-6}	10^{-4}

The portion of the theoretical electron density profiles (fig. 1) below 70 km illustrates the effect of cosmic radiation. The observed electron densities between 65 km and 70 km are not in accord with the theoretical prediction of a decrease in electron density below 70 km. However, in this region of the atmosphere the possible reactions between the various neutral constituents, species of ion and electrons are quite numerous and very few of the rates of these reactions are known even to within an order of magnitude.

Possible causes of the minimum at 83.5 km include a decrease in the concentration of nitric oxide in the region of decreasing temperature, and the attachment of electrons to dust. The later hypothesis deserves more attention in view of the detection of dust in the vicinity of the mesopause by Soberman, Witt, and Hemenway.[6]

CONCLUSIONS

A measurement has been made at middle latitudes of electron density between 65 and 90 km. The D region was characterized by a layer of electron density which increases from 115±35 electrons/cm³ at 70.5 km to 700±50 electrons/cm³ at 82 km. These observations were in far better agreement with a theoretical model based upon the ionization of nitric oxide by Lyman alpha rather than with a model based upon ionization by 2–8A X-rays. The nitric oxide density assumed was of the order of 10^{10} the total particle density, and the dissociative recombination coefficient used in the calculation was 4×10^{-8} cm³/sec. A decrease in the electron density has been established in the vicinity of the mesopause.

REFERENCES

1. KANE, J. A., "Re-Evaluation of Ionospheric Electron Densities and Collision Frequencies Derived from Rocket Measurements of Refractive Index and Attenuation," *J. Atmos. Terrest. Phys.*, **23**, 338–347, December 1961.
2. NICOLET, M., and AIKIN, A. C., "The Formation of the D Region of the Ionosphere," *J. Geophys. Res.*, **65**(5): 1469–1483, May 1960.
3. KASNER, W. H., ROGERS, W. A., and BIONDI, M. A., "Electron-Ion Recombination Coefficients in Nitrogen and in Oxygen," *Phys. Rev. Letters*, **7**(8): 321–323, October 15, 1961.
4. CHANIN, L. M., PHELPS, A. V., and BIONDI, M. A., "Measurements of the Attachment of Low-Energy Electrons to Oxygen Molecules," *Phys. Rev.*, **128**(1): 219–230, October 1, 1962.
5. BURCH, D. S., SMITH, S. J., and BRANSCOMB, L. M., "Photodetachment of O_2," *Phys. Rev.*, **112**(1): 171–175, October 1, 1958.
6. WITT, G., HEMENWAY, C. L., and SOBERMAN, R. K., "The Collection and Analysis of Particles from the Mesopause," presented at the IV International COSPAR Meeting, Warsaw, Poland, 1963.
7. POUNDS, K. A. (Private communication).

THE EFFECT OF AURORAL BREMSSTRAHLUNG ON THE LOWER IONOSPHERE

A. C. AIKIN AND E. J. MAIER

Goddard Space Flight Center

Ionization of the lower ionosphere by electrons from a type I aurora is discussed. The primary electrons produce ionization in the altitude range where in quiet conditions there is normally no ionization at night, but where the ionization during the day results from solar Lyman α and X-ray radiation. The bremsstrahlung resulting from the assumed flux of auroral electrons produces significant ionization at low altitudes (50–70 km) where normally cosmic rays are the only source of ionization. For certain values of the relevant parameters, predicted electron and ion densities, are given here for comparison with the charged particle profiles calculated for the case of no auroral flux.

INTRODUCTION

Rocket measurements by McIlwain[1] have demonstrated that for certain types of auroral events a major portion of the luminosity can be accounted for by the action of electrons with energies less than 10 kev. This energy is, however, insufficient to allow the particles to penetrate the D region. The enhanced ionization of the lower ionosphere during an auroral event is in fact caused by the energy loss of electrons whose energy distribution falls between 10 and 200 kev. In addition to direct ionization, which occurs above 70 km, ion-pair production can also take place as a result of the indirect process of energy loss by radiation.

Meredith, Gottlieb, and Van Allen[2] first demonstrated that there was a photon flux in the energy range 10 kev to 100 kev resulting from bremsstrahlung production by auroral zone electrons. This radiation can penetrate well below 70 km and leads to enhanced ionization of the lower D region, which some investigators have called the C region. That enhancement will be the topic of this discussion. In addition to ionization by bremsstrahlung it will be necessary to include effects of background cosmic rays which are the only source of ionization in this height range of the earth's atmosphere in the absence of solar flares.

The effect of auroral protons will be neglected. For this type of event protons have in general neither the flux nor the energy spectrum required to change the ion and electron densities appreciably. Loss processes considered important will be discussed and electron density profiles for various conditions will be derived.

DIFFERENTIAL ENERGY SPECTRUM OF THE PARTICLES

In order to calculate the bremsstrahlung it is necessary to know the number and energy distribution of the auroral electrons at any atmospheric depth x. It will be assumed that the incident flux of particles is isotropic over a solid angle of 2π steradians and enters the atmosphere at a pitch angle θ with respect to a vertical magnetic field. McIlwain[1] has made a rocket measurement of the electron spectrum during a type I aurora. His result can be approximated by a differential energy spectrum of the form

$$i_0 = \beta E^{-n} \qquad (1)$$

where E is the kinetic energy of the particles, $\beta = 1.6 \times 10^{12}$ particles/cm^2-sec-kev-ster, and $n = 5.2$.

As the electrons penetrate to greater depths, their interaction with the atmosphere causes a change in the form of the particle spectrum. This can be represented by an exponential attenuation of the incident spectrum such that at any level x in the atmosphere the differential energy spectrum is given by

$$i(E,x) = 2\pi\int_{\theta=0}^{\pi/2} i_0(E)\exp\left(\frac{-x}{bE^m\cos\theta}\right)\sin\theta\,d\theta \tag{2}$$

where integration has been indicated over the polar angle and includes the pitch angle dependence. Following Maeda[3] $b=3.15\times10^{-7}$ gm/cm^2 and $m=2.2$. The values of b and m are valid for $5<E<300$ kev.

PRODUCTION OF BREMSSTRAHLUNG PHOTONS BY AURORAL ELECTRONS

The cross section per atom for production of photons of energy $h\nu$ by an electron of energy E is[4]

$$\phi(E,h\nu)\,d\nu = \frac{8}{3}\alpha\rho_0^2 Z^2mc^2\frac{1}{\nu E}\ln\frac{\left(\sqrt{E}+\sqrt{E-h\nu}\right)^2}{h\nu}\,d\nu \tag{3}$$

for energies much less than 510 kev. Here α represents the fine structure constant, ρ_0 the classical radius of the electron and Z the charge of the nucleus. The differential energy spectrum of photons $dQ(h\nu)$ emitted at an atmospheric depth x in a thickness dx is

$$dQ(h\nu)d\nu = k\int_{E=h\nu}^{\infty}\phi(E,h\nu)\,i(E,x)\,dE\,dx\,d\nu \tag{4}$$

where the electron omnidirectional differential energy spectrum $i(E, x)$ described previously must be employed, and k is the number of atoms per gram. Note that only electrons of energy $E\geq h\nu$ contribute to the photon flux at energy $h\nu$. Thus the integration over E is carried out from $h\nu$ to ∞.

The total production of bremsstrahlung is then the integral over atmospheric depth of equation 4:

$$Q(h\nu)\,d\nu = k\int_{x=0}^{0.05}\int_{E=h\nu}^{\infty}\phi(E,h\nu)\,i(E,x)\,dx\,dE\,d\nu \tag{5}$$

The value $x=0.05$ gm/cm^2 yields an attenuation of at least e^{-1} for 200 kev electrons. This, combined with the relative sparsity (E^{-n}) of high energy particles allows us to set $x=0.05$ gm/cm^2 as the upper limit for the integration over the source volume. Evaluating equation 2 over x gives

$$\int_0^{0.05} i(E,x)\,dx = 2\pi\beta b E^{m-n}\,. \tag{6}$$

Then

$$Q(h\nu)\,d\nu = K\beta\int_{E=h\nu}^{E=\infty}\frac{E^{m-n-1}}{\nu}\ln\frac{\left(\sqrt{E}+\sqrt{E-h\nu}\right)^2}{h\nu}\,dE\,d\nu \tag{7}$$

where

$$K\beta = 2\pi\beta bk\frac{8}{3}\alpha\rho_0^2Z^2mc^2$$

$$= 1.62\times10^{-6}\beta\,.$$

Integrating by parts, we have

$$Q(h\nu)\,d\nu = K\beta\frac{d\nu}{\nu}\int_{E=h\nu}^{E=\infty}E^{m-n-1}\ln\frac{\left(\sqrt{E}+\sqrt{E-h\nu}\right)^2}{h\nu}\,dE \tag{8}$$

$$= K\beta\frac{d\nu}{\nu}\frac{16}{45}\frac{1}{(h\nu)^3}$$

The number of photons in the interval $h\nu_1$ to $h\nu_2$ is

$$Q(h\nu) = \frac{16}{45} K\beta \int_{h\nu_1}^{h\nu_2} \frac{d(h\nu)}{(h\nu)^4} = 0.19\beta \times 10^{-6} \times \frac{1}{(h\nu)^3} \Bigg|_{h\nu_2}^{h\nu_1} . \quad (9)$$

Table 1 gives the number of photons/cm²-sec-kev as a function of bremsstrahlung energy.

Anderson and Enemark[5] have derived an expression for the number of photons per unit energy produced by electrons having differential energy spectra of the form BE^{-n} and Ae^{-bE}. Their calculation differs from that of the present authors in the use of the range-energy relationship $R = E/2000$ gm/cm² to describe the stopping of electrons and in the use of a single bremsstrahlung radiation length, l_R, for the production of X-rays. They obtained the result (for the case of a power law electron spectrum):

$$Q(h\nu_1, h\nu_2) = \frac{B}{2000\, l_R\, (n-1)(n-2)^2} \left[\frac{1}{(h\nu_1)^{n-2}} - \frac{1}{(h\nu_2)^{n-2}} \right] . \quad (10)$$

For $l_R = 100$ gm/cm² this reduces to

$$Q = 0.12\, B \times 10^{-6} \left[\frac{1}{(h\nu_1)^{3.2}} - \frac{1}{(h\nu_2)^{3.2}} \right] ,$$

where B is in electrons/cm²-sec-kev. To compare this to equation 9 we must replace B by $2\pi B$, and we have[5]

$$Q = 0.75\, \beta \times 10^{-6} \left[\frac{1}{(h\nu_1)^{3.2}} - \frac{1}{(h\nu_2)^{3.2}} \right]$$

For the given input spectrum the present calculation thus predicts only about $\frac{1}{3}$ as many photons.

ATMOSPHERIC IONIZATION BY BREMSSTRAHLUNG

Having obtained the radiation flux incident at an atmospheric depth of 0.05 gm/cm², we must now consider the absorption of this flux as it penetrates into the lower mesosphere and stratosphere. For the energy range of interest, 10 kev to 100 kev, the mechanisms for photon absorption are (1) the photoelectric effect and (2) Compton scattering. In an energetic photoelectric interaction essentially all the photon's energy is transferred to the liberated electron. This electron can then produce secondary ionization at the rate of 1 ion per 32 electron volts of energy. The Compton effect, however, imparts a significant fraction of the incident energy to the scattered photon so that in a single interaction only a fraction of the energy is available to produce secondary ionization. Evans[6] has tabulated the product, σ_a, of the Compton scattering cross section σ_T and the average fraction of energy transferred to the scattered electron for various energies. This will be employed to obtain the efficiency of ionization by the Compton absorption process.

The number of ion pairs/cm³-sec produced by bremsstrahlung at any altitude below $x = 0.05$ gm/cm² is expressed by

$$q = n(h)\, (\sigma_p + \sigma_a)\, \frac{h\nu}{.032}\, Q_\infty(h\nu)\, e^{-\tau} \quad (11)$$

Here $n(h)$ is the number density of molecules at a height, $Q_\infty(h\nu)$ is the photon flux given in table 1, and σ_p is the cross section for photoelectric absorption as calculated by Gradstein.[7] The total absorption of the bremsstrahlung photons is repre-

TABLE 1.—*Auroral bremsstrahlung production for various photon energies*

Mean Photon Energy (kev)	$\left(\frac{\text{photons}}{\text{cm}^2\text{-sec-kev}}\right)$
10	113
15	20
20	5.0
30	1.25
40	.40
50	.15
60	.066
80	.024

sented by T which is expressed as

$$T = (\sigma_p + \sigma_a)\ \Delta n(h) H \quad , \qquad (12)$$

where $\Delta n(h) H$ is the number of molecules between the height where the photon flux is unattenuated and any lower height h.

The foregoing expressions for q and T neglect the effect of the energy released from Compton interactions in the form of scattered photons. This energy is considered as neither locally absorbed nor neglected, but as transferred to the lower altitude region. An upper limit for the ionization produced in the lower altitude region has thus been computed. To assign a lower limit to the ionization vs. altitude, the calculation has been repeated with σ_T, the Compton scattering cross section, in place of σ_a in equations 11 and 12. This procedure regards all the energy of an electron involved in a Compton interaction as locally absorbed, with none transferred to lower altitudes. The results of this calculation were not significantly different from those involving σ_a, and will not be presented here.

Figure 1 illustrates the q's for different photon energies resulting from the use of equation 11. It can be seen that most of the bremsstrahlung ion-pair production occurs for photons whose energy is less than 20 kev.

In order to calculate the total ionization below 100 km it is necessary to include the effect of cosmic rays and auroral electrons. Nicolet and Aikin[8] have shown that the effect of cosmic rays can be expressed as $q_{CR} = 10^{-17}\ n(h)$. Ionization energy loss by the primary auroral electrons is given by

$$q_\beta = \frac{p(h)}{0.032} \int_{E_{min.}}^{E_{max.}} K_e(E)\ i(E,h)\ dE \qquad (13)$$

where $p(h)$ is the atmospheric density at height h and $K_e(E)$ is the ionization energy loss formula for electrons. The energy loss expression corresponding to the exponential electron absorption which has been used is

$$K_e(E) = \frac{1}{3.5 \times 10^{-7}\ E^{1.2}} \frac{\text{kev}}{\text{gm/cm}^2} \qquad (14)$$

for E expressed in kev.

The integration involved in q_β can be carried out analytically for incident spectra of the form

$$i_0(E) = \beta E^{-n} \quad \text{where} \quad n = 4.2 \text{ or } 6.2$$

For the case $n = 5.2$, the values for q_β were obtained by a graphical interpolation between the two analytic solutions where the two theoretical expressions had been normalized to incident spectra containing the same number of electrons at energies above 10 kev.

A comparison between the rates of ionization by cosmic rays, auroral electrons and bremsstrahlung is shown in figure 2. Below 50 km bremsstrahlung plays practically no part in the ionization of the atmosphere for the number density and energy distribution of the particles considered in this work.

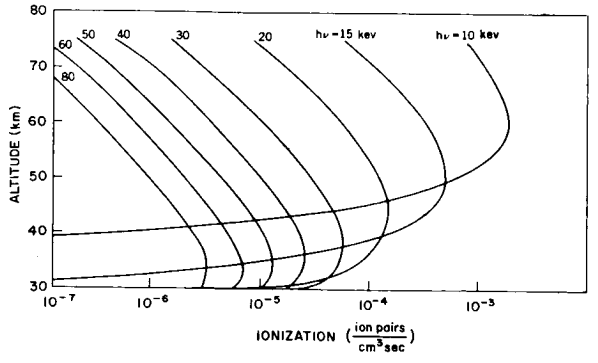


FIGURE 1.—Ion-pair production functions due to bremsstrahlung, as functions of photon energy and altitude.

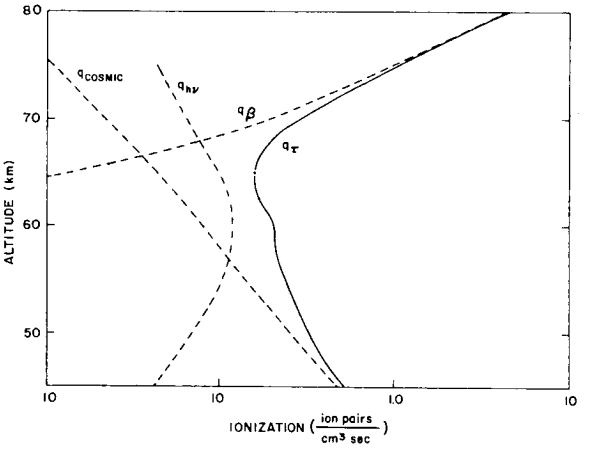


FIGURE 2.—Ion-pair production functions for the various ionization mechanisms, as functions of altitude.

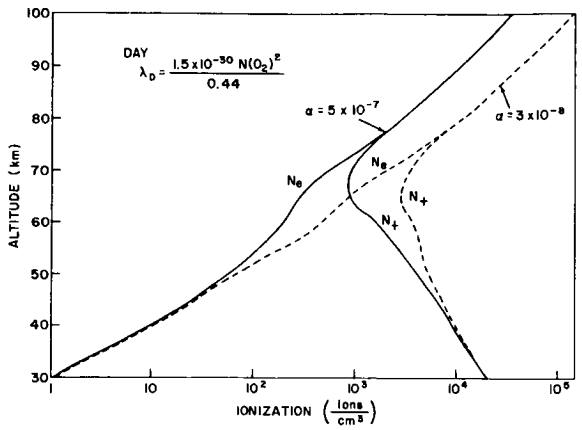


FIGURE 3.—Daytime ionization of the mesosphere due to auroral electrons.

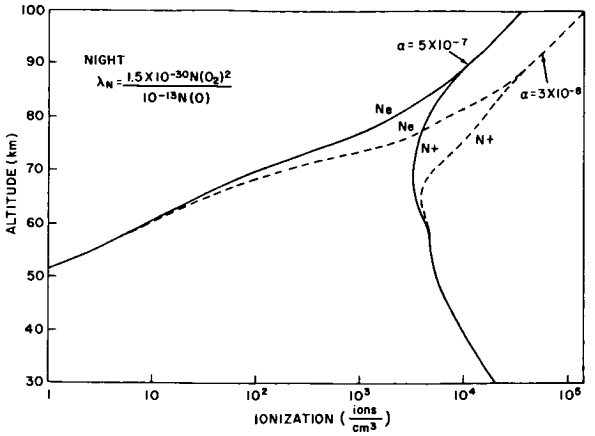


FIGURE 4.—Nighttime ionization of the mesosphere due to auroral electrons.

In order to calculate the electron density it is necessary to sum the q's described above and take into account loss processes, which will be discussed in the next section.

DISCUSSION OF LOSS PROCESSES

Electron and ion densities in the ionosphere are the result of both ionization and recombination. It will be assumed that electrons are lost either by attachment to form negative ions or by dissociative recombination with positive ions. For loss by dissociative recombination, rate coefficients of 3×10^{-8} cm^3/sec Aikin[8,9] and 5×10^{-7} cm^3/sec (reference 10) have been chosen to exhibit the effect of this process on the distribution.

Processes have been included for the formation and loss of negative ions as discussed by Nicolet and Aikin.[8] O_2^- is considered to be the predominant negative ion. It is formed by the process of three-body attachment for which the rate is

$$a = 1.5\times10^{-30}\ \text{cm}^6/\text{sec}.$$

During the day, photodetachment is operative and

$$d = 0.44\ \text{sec}^{-1}.$$

At night associative detachment of the form

$$O_2^- + O \rightarrow O_3 + e$$

is included. This process has been assigned a rate coefficient of 10^{-13} cm^3/sec. Associative detachment may be particularly important in auroral events, since there is considerable dissociation of molecular oxygen by auroral electrons as discussed by Maeda.[3]

Ionic recombination of the form

$$XW^- + YZ^+ \rightarrow XW + YZ$$

is operative and the rate coefficient α_1 is taken to be 10^{-8} cm^3/sec. At low altitudes the Thompson three-body process must be included and

$$\alpha_T \approx 10^{-8} p$$

where k is in mm Hg.

For a production function q the equations governing the distribution of ionization under equilibrium conditions are

$$\left.\begin{aligned} N^+N_e &= \frac{q}{\alpha_D + \lambda(\alpha_1 + \alpha_T)} \\ N^+ &= (1+\lambda)\,N_e = N^- + N_e \\ \lambda_{day} &= \frac{N^-}{N_e} = \frac{1.5\times10^{-30}\,n(O_2)^2}{0.44} \\ \lambda_{night} &= \frac{1.5\times10^{-30}\,n(O_2)^2}{10^{-13}\,n(O)} \end{aligned}\right\} \qquad (15)$$

Figures 3 and 4 give the electron and ion density distributions for an auroral event where the incident electron flux is that described previously. Both day and night conditions are described.

CONCLUSIONS

It has been shown that bremsstrahlung resulting from auroral electrons can be an important source of ionization in the lower ionosphere. The effect of the bremsstrahlung is to extend the region of ionization of the primary electrons to altitudes where normally cosmic rays are the only source of ionization. When the energy spectrum of the particles can be represented as $1.6\times10^{12}\ E^{-5.2}$ particles/cm^2-sec-kev-ster, the ionization due to bremsstrahlung plays a significant role from 50 km to 70 km.

REFERENCES

1. McIlwain, C. E., "Direct Measurement of Particles Producing Visible Auroras," *J. Geophys. Res.*, **65**(9): 2727–2747, September 1960.
2. Meredith, L. H., Gottlieb, M. B., and Van Allen, J. A., "Direct Detection of Soft Radiation Above 50 Kilometers in the Auroral Zone," *Phys. Rev.*, **97**(1): 201–205, January 1, 1955.
3. Maeda, K., "Auroral Dissociation of Molecular Oxygen in the Polar Mesosphere," *J. Geophys. Res.*, **68**(1): 185–197, January 1, 1963.
4. Heitler, W., "The Quantum Theory of Radiation," 3d ed., Oxford: Clarendon Press, 1954.
5. Anderson, K. A., and Enemark, D. C., "Balloon Observations of X-rays in the Auroral Zone, II," *J. Geophys. Res.*, **65**(11): 3521–3538, November 1960.
6. Evans, R. D., "Compton Effect," in: *Handbuch der Physik*, vol. XXXIV, pp. 218–298, Berlin: Springer-Verlag, 1958.
7. Gradstein, G. W., "X-ray Attenuation Coefficients from 10 Kev to 100 Mev," Nat. Bur. Standards Circular 583, April 30, 1957.
8. Nicolet, M., and Aikin, A. C., "The Formation of the D Region of the Ionosphere," *J. Geophys. Res.*, **65**(5): 1469–1483, May 1960.
9. Smith, L. G., "Rocket Measurements of Electron Density and Temperature in the Nighttime Ionosphere," Geophysics Corp. of America, Bedford, Mass., GCA TR-62-1-N, January 1962.
10. Kasner, W. H., Rogers, W. A., and Biondi, M. A., "Electron-Ion Recombination Coefficients in Nitrogen and in Oxygen," *Phys. Rev. Letters*, **7**(8): 321–323, October 15, 1961.

SATELLITE OBSERVATIONS OF COSMIC RADIO NOISE BELOW 10 Mc/s AT LOW LATITUDES

J. K. ALEXANDER

Goddard Space Flight Center

Measurements of cosmic radio noise at 3, 4, 8, and 9 Mc/s by the Alouette satellite for transits near Hawaii in November 1962, are discussed. The temperature spectral index for radiation averaged over a hemisphere centered at declination $= +10°$ and right ascension $= 20^h$ is found to be -1.8 ± 0.3 in this frequency range. A correction for synchrotron emission from the artificial radiation belt is assumed, and the resultant cosmic noise spectrum is found to show considerable flattening below 10 Mc/s in agreement with other measurements.

INTRODUCTION

The Alouette satellite was launched in September 1962, as a joint effort of the Canadian Government and the National Aeronautics and Space Administration to study the electron density distribution in the upper ionosphere by the swept radio-frequency sounding technique. In addition to the ionospheric investigations which have been discussed elsewhere (Chapman, 1962; Blumle, Fitzenreiter, and Jackson, 1963) the sounder receiver may be used to obtain an approximate measure of cosmic radio noise intensities in the 1 Mc/s to 12 Mc/s frequency range. This report presents the results of three sets of Alouette cosmic noise measurements made over Hawaii in November 1962, when the region of the galactic plane between the center and Cygnus was above the local meridian.

SATELLITE RECEIVER AND ANTENNA SYSTEM

The Alouette sounder is a swept-frequency pulsed sounder which covers the frequency range 0.5 to 12.0 Mc/s. A 100 μs pulse is repeated at a rate of 67 pulses/second with a frequency sweep rate of approximately 1 Mc/s/s. A pulse timing circuit provides for the sounder receiver to be inoperative during transmission. The receiver employs a wide-band preamplifier and two stages of frequency conversion and has an 0.3 Mc/s output bandwidth. Telemetry voltage for cosmic noise measurements over a 40 db dynamic range is provided by the AGC voltage used to control the gain of the first IF amplifier. Sufficient integration occurs to smooth the pulse echo reflected from the ionosphere so that it does not appear on the cosmic noise record. The frequency sweep is repeated every 18 seconds so that between two successive measurements at a given frequency the satellite has moved approximately 1° in declination.

The sounder antennas are a pair of orthogonal dipoles, one 150 feet tip-to-tip, the other 75 feet tip-to-tip, and are connected through parallel matching networks to a transmit-receive switch which provides isolation in excess of 50 db. Since the matching networks are designed to provide a reasonable match over a wide band, the 150 foot dipole is effective over the range 1.6 Mc/s to 4.5 Mc/s and the 75 foot dipole is effective over the range 4.5 Mc/s to 11.5 Mc/s with a loss of never more than 11 db.

Since there is no inflight calibration of the Alouette receiver parameters, the cosmic noise measurements are based upon prelaunch calibration of the system gain and impedance and calculated response characteristics. Since the mismatch loss is severe below 1.6 Mc/s and above

11.5 Mc/s and since there is some uncertainty about the system behavior in the transition region near 4.5 Mc/s, only data in the frequency range 2 to 4 Mc/s and 8 to 10 Mc/s have been used in this preliminary investigation. The satellite is in a nearly circular orbit at an altitude of 1000 km where the local plasma frequency is generally near 1 Mc/s, and, therefore, measurements below about 3 Mc/s are subject to an additional uncertainty due to plasma effects.

ANTENNA TEMPERATURES

The Alouette orbit is inclined to the equator by 80° so that at low latitudes the trajectory is one of nearly constant hour angle. In November 1962, the orbital plane was so oriented that the satellite made a series of daily transits near the South Point, Hawaii, tracking station in the late afternoon when the ionospheric critical frequency was above 10 Mc/s. Hence the ionosphere acted as a shield to prevent interference from radio noise of terrestrial origin which has been found to be a considerable problem at times when the satellite is above a region of relatively low ionization density.

Figure 1 gives a logarithmic plot of antenna temperature as a function of frequency for measurements at 3, 4, 8, and 9 Mc/s taken on November 10, 12, and 21, 1962. In each case the satellite was in the region 0° to +25° declination and right ascension approximately 20 hours. Individual measurements were averaged over the full time the satellite was in range of the tracking station to remove secular variations and any effects due to satellite spin (~1.3 rpm), and then an average of the antenna temperatures obtained for the three transits was taken to construct the spectrum in figure 1. Since the cosmic noise temperature varies with frequency according to the power law $T \propto \nu^{\alpha}$, the slope of the spectrum on a log-log plot is equal to the temperature spectral index, α. By fitting the best straight line to the points thus obtained we may arrive at an estimate of the temperature spectral index, which is found to be $\alpha = -1.9$. Due to the fact that the receiver response is down (by an amount which is not precisely known) at 9 Mc/s, the 8 Mc/s observation must be weighted in accordance with the response characteristics to select the best straight line.

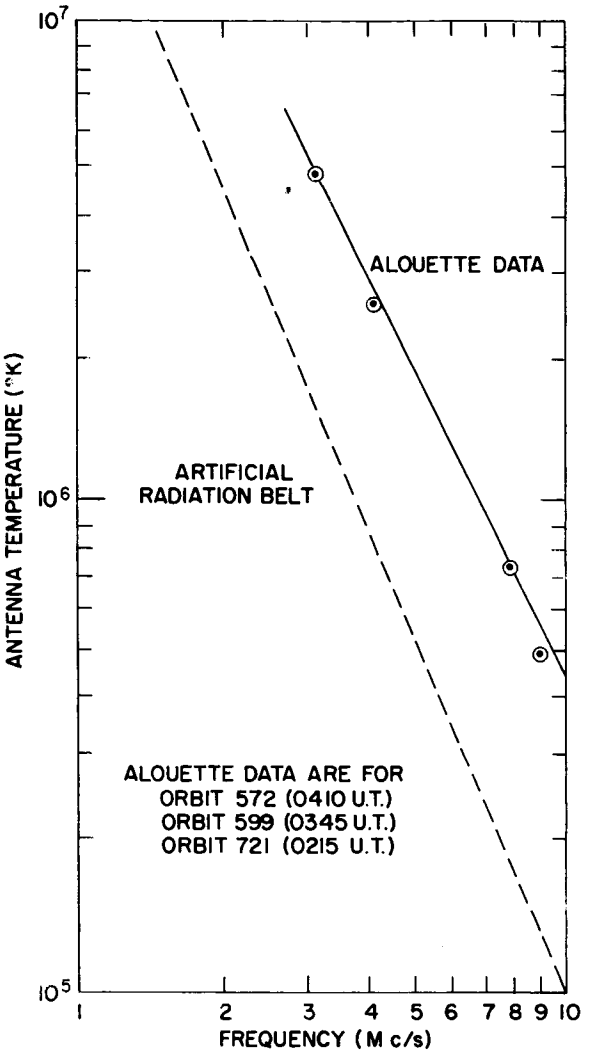


FIGURE 1.—Antenna temperatures as a function of frequency for ALOUETTE cosmic noise measurements in the region centered at declination = +10° and right ascension = 20^h. The dotted curve gives the calculated contribution from the artificial radiation belt.

DISCUSSION

In order to relate the observed antenna temperatures to the actual sky brightness at the satellite we make some rather general assumptions which, nevertheless, will allow us to obtain qualitatively good results. The satellite receives noise incident directly from the hemisphere about the zenith and noise reflected upward from the ionosphere below. The antenna temperatures derived by averaging over a number of satellite rotations

are therefore average values for the full sphere about the receiver.

It is important to note that the satellite receives noise from two major sources—cosmic radio radiation and synchrotron emission from the artificial radiation belt. Nakada (1963) has calculated the temperature spectrum for the fission electrons 2 weeks after the Starfish explosion, and the values derived from his spectrum are also shown in figure 1. A correction has been made for the decay of the belt assuming a 60-day half-life and for the solid angle subtended by the radiation belt noise source which was assumed to be 2π steradians. In the worst case, the belt subtends half the antenna beam.

An estimate of the cosmic noise spectrum may be obtained by subtracting the radiation belt temperatures (divided by 2) from the values obtained by the Alouette measurements, and this is given in figure 2. If the noise from the artificial belt is distributed over a solid angle which is less than 2π sterad then the correction to be subtracted will be less, and the resultant temperatures in figure 2 should be correspondingly higher. The results obtained from the ground by Ellis et al. (1962) for a similar region of the galaxy are higher; however the two sets of data are found to be in good agreement.

Among the possible sources of error in absolute measurement of antenna temperature we must include the following:

(1) Telemetry recording and readout $\pm 5\%$;
(2) Interpretation of AGC calibration curves ± 1 db;
(3) Measurement of receiver input impedance $\pm \frac{1}{2}$ db;
(4) Receiver gain variations—unknown;
(5) Uncertainties in the matching network ± 1 db.

The net uncertainty, therefore, is estimated to be on the order of $\pm 50\%$. The possible error in the spectral index, however, is not thought to be as high since the probable effects of uncertainties in the calibration and receiver instabilities would be to move the curve as a whole rather than to act only at one end of the frequency range. It is important to note that the numbers given here are only rough estimates of the possible error since this is not meant to be a high precision experiment. No attempt has been made to correct for detuning of the antenna by the ambient plasma, but these effects are not likely to be severe for the particular set of observations under consideration. The main effect of the radiation belt noise is to make the total observed noise spectrum steeper than it would appear to be with cosmic noise as the only source, and hence the observed flattening in the spectrum below 10 Mc/s appears to be real.

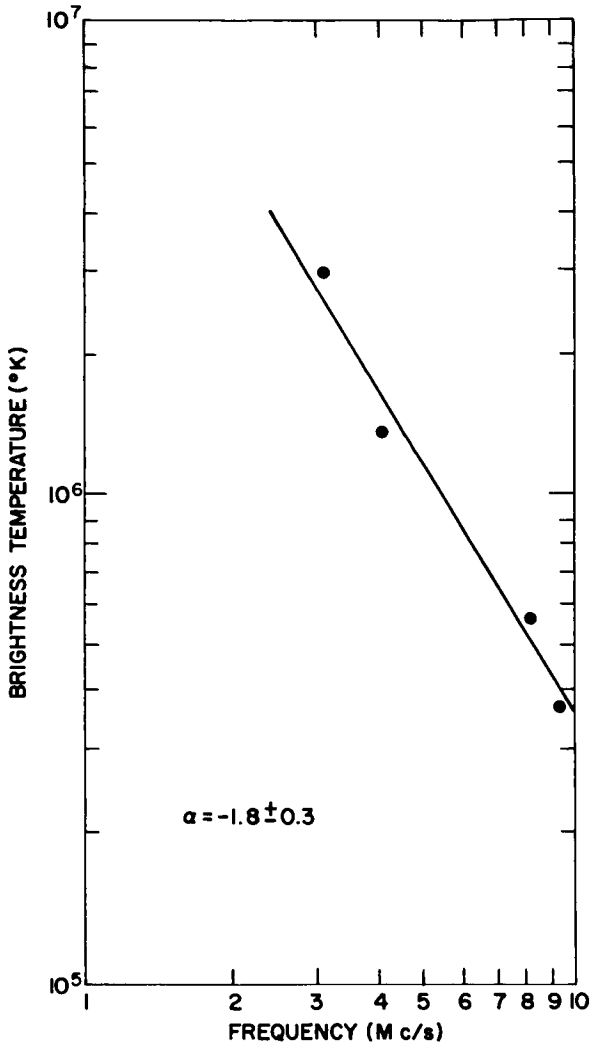


FIGURE 2.—Resultant cosmic noise spectrum after correcting for the artificial radiation belt. The temperatures should be lower limits since the correction is for the maximum likely contribution from the belt.

CONCLUSIONS

Alouette observations of cosmic radio noise from the hemisphere centered at about 10° declination and 20 hours right ascension have been found to be consistent with a temperature spectral index less than -2 in the frequency range 3 to 9 Mc/s. We must emphasize that the measurements represent an average over this region which includes radiation originating near the galactic plane, the halo, and outside our galaxy. The apparent flattening in the spectrum and low noise temperatures may be due, in part, to increased absorption in the galactic plane. By making a correction for the noise spectrum calculated for the artificial radiation belt, the data are found to be in qualitative agreement with the observations from the ground by Ellis and coworkers. The most important needs in the future are improved flux measurements below 5 Mc/s based upon an accurate in-flight calibration of the receiver and antenna system and a clear understanding of the behavior of the antenna in the ionospheric plasma. The next step is the development of a space radio telescope having enough directivity to permit investigation of particular regions of smaller angular extent.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the help of Dr. C. A. Franklin and Dr. T. R. Hartz of the Defence Research Telecommunication Establishment who supplied the Alouette calibration data and Dr. M. P. Nakada of Goddard Space Flight Center who offered many useful suggestions concerning the artificial radiation belt. The author is also indebted to Dr. R. G. Stone and Dr. G. Westerhout for numerous constructive discussions.

REFERENCES

Blumle, L. J., Fitzenreiter, R. J., and Jackson, J. E., "The NASA Topside Sounder Program," GSFC Rpt. X-615-63-9 (1963).

Chapman, J. H., "Topside Sounding of the Ionosphere," *Advances in the Astronautical Sciences*, **12** (1962).

Ellis, G. R. A., Waterworth, M. D., and Bessell, M., "Spectrum of the Galactic Radio Emission between 10 Mc/s and 1.5 Mc/s," *Nature*, **196**, 1079 (1962).

Nakada, M. P., "Synchrotron Radiation Calculations for the Artificial Radiation Belt," *J. Geophys. Res.*, in press (1963).

CONSTITUTION OF THE ATMOSPHERE AT MAGNETOSPHERIC LEVELS

S. J. BAUER

Goddard Space Flight Center

COMPOSITION

Prior to 1961 it was generally accepted that the main constituent of the atmosphere above 300 km was atomic oxygen up to an altitude of about 1000 km and that at higher altitudes hydrogen would predominate. Direct experimental evidence for the presence of neutral hydrogen in the outer atmosphere has come from the high resolution Lyman α spectrum obtained by Purcell and Tousey [1960]. From the absorption core of their Lyman α spectrum, the total content of hydrogen above the altitude of the rocket measurement has been determined. Johnson [1961] has interpreted this hydrogen content to be distributed in the form of a "geocorona." The problem of interpretation of the Lyman α observations in terms of hydrogen and its location with respect to the earth has recently been reviewed in detail by Donahue [1962]. The computation of the distribution of neutral hydrogen is complicated by the fact that at magnetospheric levels "exospheric" conditions prevail, i.e., the mean free path of neutral hydrogen is greater than the local scale height. Because of the escape of particles with high velocities, the velocity distribution is, strictly speaking, non-Maxwellian and the distribution of density with height does not follow the simple hydrostatic equation, but has to be computed by taking into account ballistic escape, ballistic reentry and bound-orbiting particles. Such computations have been made by Öpik and Singer [1961] and by Johnson [1961]. Out to a geocentric distance of about two earth radii, a simple hydrostatic distribution however, still represents a reasonably good approximation.

Photoionization and/or charge exchange with oxygen ions of the neutral hydrogen of the geocorona lead to the protons constituting the "protonosphere." The ions at magnetospheric levels are distributed according to a diffusive equilibrium distribution, i.e., hydrostatically supported but constrained by the earth's magnetic field. For ions the mean free path is short enough so that the concept of an ion-exosphere is not applicable [Johnson, 1962]. While the diffusive equilibrium distribution of an ionic species was generally considered to be governed by a scale height twice that of the corresponding neutral species, it was pointed out by Mange [1960, 1961] that this concept is not justified for a minor ion in the presence of other ions. The electric field, which is set up to prevent further charge separation of electrons and ions diffusing under gravity, and which is proportional to the mean mass of the (singly charged) positive ions, causes the density of minor light ions first to *increase* with altitude until it becomes predominant, after which it shows the usual exponential decrease with altitude according to a scale height approaching twice that of the corresponding neutral constituent. Subsequently to Mange's independent derivation of the equation governing the distribution of singly charged ions in the presence of other singly charged ions, it was realized that the same problem had been solved before by Dungey [1955], and originally as long as four decades ago [Pannekoek, 1922; Eddington, 1926]. The distribution of protons in the protonosphere according to this concept, and their relation and coupling to the ionospheric F-region where oxygen ions predominate, have been discussed in great detail by Hanson and Ortenburger [1961].

In 1961, Nicolet [1961] suggested that neutral and ionized helium should be an important constituent in the upper atmosphere. He pointed out that the presence of neutral helium would provide a sensible solution to the problem of high atmospheric densities at 1600 km deduced from the drag of the Echo I balloon satellite. The first experimental evidence, from charged particle observations, for the importance of helium in the upper atmosphere was presented before a session of U.S. Commission 4 at the Fall-URSI Meeting in Austin, Texas, 1961 [Hanson, 1962; Bourdeau, Whipple, Donley, and Bauer, 1962]. Hanson [1962] inferred the presence of helium ions from an ion density profile obtained by Hale (1961) with a Scout rocket. Hanson concluded from the analysis of this data that the concentration of He^+ was in agreement with Nicolet's estimates and that the layer where He^+ was predominant extended over about 2000 km, from 1200 to 3400 km, at a time when the atmospheric temperature corresponded to 1600°K. He also suggested that He^+ is lost by an ion-atom interchange process involving molecular nitrogen, and that no large diuranl change in the helium ion concentration should occur. More recently Bates and Patterson [1962] have shown that the loss process for He^+ with N_2 is not possible, but that the one involving O_2 seems to be responsible for the loss of He^+.

The first direct experimental evidence for the presence of He^+ was provided by the ion retarding potential measurement on Explorer VIII [Bourdeau, Whipple, Donley, and Bauer, 1962]. This measurement showed a ratio of $He^+/O^+ = 1.3 \pm 0.3$ at an altitude of 1630 km, where the simultaneously measured electron temperature was 1750 ± 200°K. Additional evidence for a transition from O^+ to He^+ rather than directly to H^+, has come from a rocket measurement of the electron density distribution [Bauer and Jackson, 1962]. The presence of He^+ and H^+ in the upper atmosphere has now also been identified directly with an RF ion spectrometer [Taylor, Brinton, and Smith, 1962]. More recently ion composition measurements on the Ariel satellite have also shown the presence of He^+ as well as a significant diurnal-variation of the altitude range where He^+ is the predominant ion [Willmore, Boyd, and Bowen, 1962]. These data, as well as a recent nighttime ion density profile obtained with a Scout rocket [Donley, 1963], indicate that at low atmospheric temperatures the layer where He^+ is predominant is only a few hundred km thick compared to a thickness of 4000 km at 1600°K determined by Hanson, with correspondingly lower altitudes of transitions from O^+ to He^+ and He^+ to H^+. Such a strong variation with temperature in the thickness of the helium ion layer in accordance with a suggestion by Bauer[1963].

The possible presence of doubly charged oxygen ions ($O^{+\,+}$) in the upper ionosphere has been suggested by Nakada and Singer [1962] on the basis of the abundance of O^+ and an adequate photoionization rate for their formation. In this connection they have investigated the distribution of multiply charged ions in an ion mixture, for which no analytical expression of the kind derived by Mange [1960] exists, and have shown by numerical integration that the concentration of $O^{+\,+}$ would also increase with altitude relative to O^+ and may, depending on loss rates, become an important ionic constituent. The presence of $O^{+\,+}$, however, has not yet been verified experimentally.

TEMPERATURE

The kinetic gas temperature of the atmosphere at magnetospheric levels has been determined from satellite drag observations assuming model distributions of the mean molecular weight. Above 300 km this temperature should be independent of altitude because of the high thermal conductivity. The kinetic gas temperature in this isothermal region has been found to vary with the solar cycle and to show short-term fluctuations which are correlated with the 10.7 cm radiation from the sun, as well as with geomagnetic activity [Jacchia, 1961; Priester, 1961; Harris and Priester, 1962]. A diurnal variation having a minimum at 4 a.m. local time and a maximum at about 2 p.m. is indicated from the satellite data.

While the kinetic gas temperature in the upper atmosphere has only been inferred from satellite drag measurements, direct measurements of the electron and ion temperatures have been made with the help of space vehicles. The vertical profiles of electron temperature up to about 400 km have been measured by means of the Langmuir probe technique [Spencer, Brace, and Carignan, 1962; Brace, 1962]. These measurements indicate

that at midlatitudes for quiet ionospheric conditions, the daytime electron temperature attains a maximum at about 230 km, approaching an isothermal behavior at altitudes above 300 km.

This measurement is in good agreement with theoretical investigations [Hanson and Johnson, 1961; Hanson, 1962; Dalgarno, McElroy, and Moffett, 1962] which show that, for solar ultraviolet as the major heat source, the only departure from temperature equilibrium would occur in the region between 200 and 400 km altitude. At higher altitudes the daytime electron temperature and the ion (and gas) temperature are expected to be equal, at least under quiet conditions. A nighttime measurement of electron temperature shows perfect isothermal behavior throughout this altitude range [Brace, 1962]. Departures from thermal equilibrium extending throughout the ionospheric F-region have been reported for disturbed ionospheric conditions at middle latitude and appear to be the rule in auroral regions [Spencer, Brace, and Carignan, 1962]. The gas temperature in the upper ionosphere has also been determined from the exponential decrement of electron and ion density profiles. These measurements show evidence of an isothermal behavior, since the scale height of the electron-ion gas is constant within a few percent over an altitude range of a few hundred kilometers. Temperatures derived from these scale heights, assuming thermal equilibrium, have been found to be in good agreement with kinetic gas temperatures expected from the correlation with solar 10.7 cm flux, thus providing indirect evidence for perature equilibrium [Bauer and Bourdeau 1962]. Direct measurements of electron temperature at magnetospheric levels have been made by means of Langmuir probes on the Explorer VIII satellite [Serbu, Bourdeau, and Donley, 1961]. These measurements also show, within their error-limits, agreement with model values of kinetic gas temperatures, except during the sunrise period when high electron temperatures seem to be prevalent. More recently, similar measurements have been made on the Ariel satellite [Willmore, Boyd, and Bowen, 1962] in the altitude region between 400 and 1200 km which indicate a latitude dependence, with midday values of 1200°K at the equator and 1600°K at a latitude of 55°N. Preliminary topside sounder satellite results are in qualitative agreement with such a latitude dependence [Knecht and Van Zandt, 1963]. The Ariel measurements also show high electron temperatures during the sunrise period while at other times the electron temperatures are equivalent to the simultaneously determined ion temperatures.

Ion temperatures and the departure from thermal equilibrium in the upper ionosphere have also been determined by ground-based magnetospheric sounders using the incoherent backscatter technique. Evans [1962] has reported measurements for a few days covering an altitude range up to 800 km, which show temperature equilibrium ($T_e = T_i$) during the night as expected, but a positive temperature gradient, and during the day departures from equilibrium in the entire region above 200 km. The latter can be interpreted either as a constant ratio T_e/T_i, with T_e and T_i showing a height gradient, or as a variable ratio T_e/T_i, with T_i=constnat and an even stronger height-dependence of T_e. The time of maximum departure from equilibrium according to his data is at noon, reaching a value of $T_e/T_i = 1.6$. This is in disagreement with satellite measurements which show high electron temperatures only during the sunrise period, as well as with incoherent backscatter measurements by Bowles, Ochs, and Green [1962], who also found the time of departure from thermal equilibrium to be only during the sunrise period and during disturbed ionospheric conditions. The present discrepancies between space-flight measurements of charged particle temperatures and those determined by means of the incoherent back-scatter technique obviously need to be resolved. It should be understood, however, that generalizations concerning the thermal properties of the upper atmosphere based as yet upon only a small number of observations may be premature.

REFERENCES

BATES, D. R., and PATTERSON, T. N. L., Helium ions in the upper atmosphere, Planet. *Space Sci.*, **9**, 599–605. (1962).

BAUER, S. J., Helium ion belt in the upper atmosphere *Nature*, **197**, 36–37 (1962).

———, and BOURDEAU, R. E., Upper atmosphere temperatures derived from charged particle observations, *J. Atmos. Sci.*, **19**, 218–225 (1962).

———, and JACKSON, J. E., Rocket measurement of the

electron density distribution in the topside ionosphere, *J. Geophys. Res.*, **67,** 1675–1677 (1962).

BOURDEAU, R. E., WHIPPLE, E. C., DONLEY, J. L., and BAUER, S. J., Experimental evidence for the presence of helium ions based on Explorer VIII satellite data, *J. Geophys. Res.*, **67,** 467–475 (1962).

BOWLES, K. L., OCHS, G. R., and GREEN, J. L., On the absolute intensity of incoherent scatter echoes from the ionosphere, *J. Res. NB S,* **66**D (Radio Prop.) 395–407 (1962).

BRACE, L. H., The dumbell electrostatic ionosphere probe: ionosphere data, Univ. of Michigan Sci. Rept. JS-3, 139 pp. (1962).

DALGARNO, A., MCELROY, M. B., and MOFFETT, R. J., Electron temperatures in the ionosphere, Geophysics Corp. of Amer. Tech. Rept. 62-11-N, 55 pp. (1962).

DONAHUE, T. M., Excitation of the Lyman α in the night sky, Space Sci. Rev., **1,** 135–153 (1962).

DONLEY, J. L., Experimental evidence for a low ion transition altitude in the upper nighttime ionosphere, *J. Geophys. Res.*, **68** (1963).

DUNGEY, J. W., The electrodynamics of the outer atmosphere, The Physics of the Ionosphere (Phys. Soc., London), (1955).

EDDINGTON, A. S., The Internal Constitution of the Stars, 272–274 (Cambridge Univ. Press) (1926).

EVANS, J. F., Diurnal temperature variation of the *F*-region, *J. Geophys. Res.*, **67,** 4914–4920 (1962).

HALE, L. C., Ionospheric measurements with a multigrid potential analyzer, *J. Geophys. Res.*, **66,** 1554 (1961).

HANSON, W. B., Upper atmosphere helium ions, *J. Geophys. Res.*, **67,** 183–188 (1962a).

———, Electron temperatures in the upper atmosphere, *Third International Space Science Symposium of COSPAR*, Washington, D.C. (1962b).

———, and JOHNSON, F. S., Electron temperatures in the ionosphere, *Memoires Soc. Roy. Sci. Liege*, Tome IV, 390–423 (1961).

———, and ORTENBURGER, I. B., The coupling between the protonosphere and the normal *F* region, *J. Geophys. Res.*, **66,** 1425–1455 (1961).

HARRIS, I., and PRIESTER, W., Theoretical models for the solar cycle variation of the upper atmosphere, *J. Geophys. Res.*, **67,** 4585–4591 (1962).

JACCHIA, L. G., A working model for the upper atmosphere, *Nature*, **192,** 1147 (1961).

JOHNSON, F. S., The distribution of hydrogen in the telluric hydrogen geocorona, *Astrophys. J.*, **133,** 701 (1961).

———, Physics of the distribution of ionized particles in the exosphere, Electron Density Profiles in the Ionosphere and Exosphere, ed. B. Maehlum, 404–413 (Pergamon Press, New York and London) (1962).

KNECHT, R. W., and VAN ZANDT, T. E., Some early results from the ionosphere topside sounder satellite, *Nature*, **197** (in press) (1963).

MANGE, P., The distribution of minor ions in electrostatic equilibrium in the high atmosphere, *J. Geophys. Res.*, **65,** 3833–3834 (1960).

———, Diffusion in the thermosphere, *Ann. Geophys.*, **17,** 277–291 (1961).

NAKADA, M. P., and SINGER, S. F., Multiply ionized oxygen in the magnetosphere, URSI-Spring Meeting, Washington, D.C. (1962).

NICOLET, M., Helium, an important constituent in the lower exosphere, *J. Geophys. Res.*, **66,** 2263–2264 (1961).

ÖPIK, E., and SINGER, S. F., Distribution of density in a planetary exosphere, II, Phys. Fluids, **4,** 221 (1961).

PANNEKOEK, A., Ionization in stellar atmospheres, Bull. Astron. Inst. Netherlands, No. 19 (1922)

PRIESTER, W., Solar activity effect and diurnal variation in the upper atmosphere, *J. Geophys. Res.*, **66,** 4143–4148 (1961).

PURCELL, J., and TOUSEY, R., The profile of solar hydrogen Lyman α, *J. Geophys. Res.*, **65,** 370–372 (1960).

SERBU, G. P., BOURDEAU, R. E., and DONLEY, J. L., Electron temperature measurements on the Explorer VIII satellite, *J. Geophys. Res.*, **66,** 4313–4315 (1961).

SPENCER, N. W., BRACE, L. H., and CARIGNAN, J. R., Electron temperature evidence for nonthermal equilibrium in the ionosphere, *J. Geophys. Res.*, **67,** 157–176 (1962).

TAYLOR, H. A., BRINTON, A. H., and SMITH, C. R., Instrumentation for atmospheric composition measurements, *Proc. 8th Aerospace Instrumentation Symposium* (Instrument Society of America, Pittsburgh) (1962).

WILLMORE, A. P., BOYD, R. L. F., and BOWEN, P. J., Some preliminary results of the plasma probe experiments on the Ariel satellite, Proc. Internat. Conf. on the Ionosphere, London, July 1962 (1962).

SOME IMPLICATIONS OF A DIRECT MEASUREMENT OF THE HYDROGEN AND HELIUM ION DISTRIBUTION IN THE UPPER ATMOSPHERE

S. J. BAUER

Goddard Space Flight Center

A recent daytime measurement of ion composition in the altitude region from 300 km to 940 km by Taylor, Brace, Brinton, and Smith (1963) allows some inferences about the physical processes governing the distribution of the light ionic constituents H^+ and He^+. Although the absolute values for the ion concentrations derived from the current-measurements of the ion mass spectrometer may be subject to some errors, useful information can nevertheless be obtained from these data. The observed proton distribution has already been used by Hanson, Patterson, and Degaonkar (1963) to deduce a value for the rate coefficient of the charge exchange reaction between hydrogen and oxygen.

By using the *ratio* of the two observed ionic constituents, $n(H^+)/n(H^+)$, it is possible to deduce additional information on the physical processes and structure parameters of the upper atmosphere.

In figure 1 the ratio $n(H^+)/n(He^+)$ based on the experimental data of Taylor et al. (1963) is plotted as a function of geopotential altitude. It is quite obvious that the altitude distribution of this ratio can be divided into three distinct regions of behavior. These can be explained as the result of the predominance of either chemical processes (charge exchange, photoionization and ion-atom interchange or radiative recombination) or diffusion processes, or their combination as shown in the following:

I. The lowermost portion of the distribution of $n(H^+)/n(He^+)$ in figure 1 can be explained by a chemical equilibrium distribution of the two ionic constituents.

It is now generally agreed that below 500 km the charge exchange reaction between hydrogen and oxygen governs the proton distribution (Hanson et al., 1963). Thus,

$$n(H^+) = \tfrac{8}{5}[n(H)/n(O)]n(O^+) \qquad (1)$$

where $n(X)$ is the number density of the constituent X, which is distributed exponentially

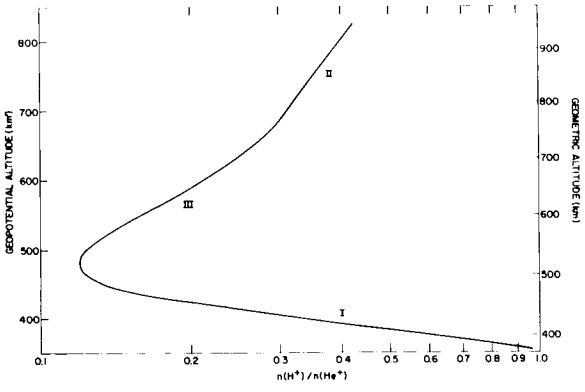


FIGURE 1.—The ratio $n(H^+)/n(He^+)$ based on the experimental data of *Taylor et al.* (1963) plotted as a function of geopotential altitude.

with geopotential altitude according to its scale height, $H(X)$.

The photochemical equilibrium distribution of He^+ is governed by

$$n(He^+) = I(He)n(He)/Kn(O_2) \qquad (2)$$

where $I(He)$ is the photoionization rate coefficient for helium (Hinteregger and Watanabe, 1962) and K is the rate coefficient for the ion-

atom interchange reaction with molecular oxygen, now considered to be responsible for the loss of helium ions (Bates and Patterson, 1962; Hanson, 1963).

It is easily shown from (1) and (2), that for the case of chemical and thermal equilibrium the ratio of the two ionic constituents follows the altitude dependence:

$$n(H^+)/n(He^+) = A \exp[-z/H(21)] \quad (3)$$

where A is a constant, z is the geopotential altitude above the chosen reference level and $H(21)$ is the scale height corresponding to an effective mass of 21 A.M.U. (It is interesting to note that the altitude dependence given in (3) also holds even if the He^+ and H^+ distributions were both controlled by photochemical equilibrium, i.e., H^+ being produced by photoionization and lost by radiative recombination, assuming O^+ to be the predominant ion). Making use of (3) for the lowermost portion of the curve in Fig. 1 leads to an upper limit for the netural gas temperature of 1200°±60°K which is in good agreement with the effective charged particle temperature $(T_e+T_i)/2 = 1235°\pm40°K$ inferred by Taylor et al. (1963) from the slope of their total ion density distribution. (The estimated neutral gas temperature for the time of this measurement, based on satellite drag observations (Jacchia and Slowey, 1963), is 1050°K).

II. The uppermost portion of the $n(H^+)/n(He^+)$ distribution corresponds to a diffusive equilibrium distribution.

It is now generally known (Mange, 1960) that (unlike the behavior of neutral constituents) the diffusive equilibrium distribution of minor ions is not independent of other ionic constituents because of the presence of an electrostatic field (proportional to the mean ionic mass and the electron and ion temperatures) which is the same for all ions in the mixture and counteracts the gravitational force.

The diffusive equilibrium distribution of an ion (X_i^+) is given by:

$$n(X_i^+) = n_0(X_i^+) \times \exp\left[-\int_{z_0}^{z}\left(\frac{1}{H(X_i^+)} - \frac{\tau}{1+\tau}\frac{m_+ g_0}{kT_i}\right)dz\right] \quad (4)$$

where $H(X_i^+) = kT_i/m_i g_0$, with k Boltzmann's constant, T_i the ion temperature, m_i the ion mass and g_0 the acceleration of gravity at the earth's surface and where $\tau = T_e/T_i$ is the ratio of electron to ion temperature, z is the geopotential altitude and $m_+ = \sum n(X_i^+)m_i/\sum n(X_i^+)$ is the mean ionic mass.

It is easily seen from (4) that the *ratio* of two ionic constituents in diffusive equilibrium depends on the ion temperature only, even if there is absence of thermal equilibrium. For the two ionic constituents H^+ and He^+ this ratio can be expressed by

$$n(H^+)/n(He^+) = B \exp[z/H(3)] \quad (5)$$

where B is a constant and $H(3)$ corresponds to the scale height for an effective mass of 3 A.M.U. Using (5), an ion temperature $T_i = 1300°\pm60°K$ is inferred from the top portion of the curve in fig. 1, in good agreement with the temperature calculated for the chemical equilibrium region and with the effective charged particle temperature inferred by Taylor et al. (1963).

III. The intermediate region in fig. 1 corresponds to a variable scale height which is a weighted mean of the scale heights for the diffusive and chemical equilibrium regions.

As the result of the foregoing discussion the ion composition data of Taylor et al. (1963), can be interpreted in the following fashion:

Helium and hydrogen ions are in chemical equilibrium up to a geometric altitude of about 500 km. From this altitude to about 750 km, both chemical and diffusion processes seem to be influencing the distribution, while above 750 km the light ionic constituents appear to follow essentially a diffusive equilibrium distribution. It should be understood that departures from a true diffusive equilibrium distribution evidenced in the intermediate region may also arise from the effect of an upward or downward flux of ions as shown by Hanson (1963). The conclusion of diffusive equilibrium for the light ions above 750 km is difficult to reconcile with the observed decrease of $n(He^+)$ at these altitudes, unless one is willing to accept that the absolute values of the light ion concentrations reported by Taylor et al. (1963), are too low, or their values of total ion density (and $n(O^+)$ which is inferred therefrom)

are too high, or that a combination of these factors is present. If the light ions were of comparable order of magnitude with $n(O^+)$ at altitudes above 750 km, $n(He^+)$ would indeed decrease with altitude as shown by Taylor et al. (1963), even though it is in diffusive equilibrium as indicated above, without reaching a second peak in its distribution. This fact can easily be verified from equation (4) by using a mean ionic mass appropriate to this situation. However, the magnitude of the adjustment of the experimental data of Taylor et al. (1963), required to produce the above effect does not appear to be justifiable on the basis of the internal consistency of their data (Taylor and Brace, private communication).

The above conclusions, if supported by future measurements, may seriously impede efforts of explaining charged particle profiles in terms of a ternary ion-mixture model in diffusive equilibrium.

It is also quite evident that even measurements of *relative* ion abundances can provide useful information on the physical processes in the upper atmosphere. In particular, they can be used to determine ion temperature, up to now a parameter most elusive to direct measurements.

REFERENCES

Bates, D. R., and Patterson, T. N. L., Helium ions in the upper atmosphere, Planet. *Space Sci.*, **9,** 599–605, 1962.

Hanson, W. B., Dynamic diffusion processes in the exosphere, *"Electron Density Profiles in the Ionosphere and Exosphere,"* Pergamon Press, 1963 (in press).

———, Patterson, T. N. L., and Degaonkar, S. S., Some deductions from a measurement of the hydrogen ion distribution in the high atmosphere, *J. Geophys. Research,* **68,** 1963 (in press).

Hinteregger, H. E., and Watanabe, K., Photoionization rates in the E and F region, 2, *J. Geophys. Research,* **67,** 3373–3392, 1962.

Jacchia, L. G., and Slowey, J., Smithsonian Institution Astrophysical Observatory Special Report No. 125, Cambridge, Mass., 1963.

Mange, P., The distribution of minor ions in electrostatic equilibrium in the high atmosphere, *J. Geophys. Research,* **65,** 3833–3834, 1960.

Taylor, H. A., Brace, L. H., Brinton, H. C., and Smith, C. R., Direct measurement of helium and hydrogen ion concentration and total ion density to an altitude of 940 km, *J. Geophys. Research,* **68,** 5339–5348, 1963.

SIMULTANEOUS ROCKET AND SATELLITE MEASUREMENTS OF THE TOPSIDE IONOSPHERE

S. J. BAUER, L. J. BLUMLE, J. L. DONLEY, R. J. FITZENREITER AND J. E. JACKSON

Goddard Space Flight Center

In view of the vast amount of data accumulating from Alouette, the Canadian Topside Sounder Satellite (Chapman, 1963), it is highly desirable to establish the accuracy of this technique by comparing the electron density profiles obtained by Alouette with similar data obtained independently. This accuracy affects the geophysical interpretation of the data as well as the possibility of using the topside sounder to calibrate ground-based techniques such as the incoherent backscatter radar. There are two possible sources of errors in the electron density profiles derived from topside sounders. The first one is associated with the method of converting the observed ionograms to true-height profiles. It has now been established that topside profiles agreeing within $\pm 5\%$ in electron density (or within ± 15 km in terms of altitude) are obtained by using various methods of analysis which differ significantly from each other in their underlying assumptions. A more serious problem is the uncertainty due to possible deviations of the propagation paths from the assumed vertical. In some cases the presence of oblique propagation is quite obvious from ionograms whose analysis yield absurdly low values for the height of the F2 maximum. However, in the general case it may be quite difficult to assess the contribution of ray bending to the observed virtual depth and the resulting true height profile. Even matching topside and bottom soundings at the same location cannot quite resolve this uncertainty because the height of the F2 maximum (which is common to both observations) is the least accurate point determined by either sounding technique.

A much more satisfactory test of the accuracy of topside sounder measurements is to obtain a topside profile at the same time and location by an independent and reliable technique. For this reason an ARGO D–4 rocket (NASA 8.14) was launched from Wallops Island, Va. on July 2, 1963, to intersect the Alouette orbit. This rocket carried two proven experiments each yielding independently the charged particle distribution in the topside ionosphere. The electron density was measured by the well-established two-frequency CW propagation technique as used in previous high-altitude rocket flights (Jackson-Bauer, 1961); in this case using 24.5 and 73.6 Mc signals. The ion density profile was measured by an ion retarding potential experiment employing a planar ion trap, similar to that used in past rocket and satellite experiments (Donley, 1963; Bourdeau, Whipple, Donley, and Bauer, 1962).

The date and time of the launching were selected to provide adequate proximity for a daytime comparison between the rocket and the satellite data. During the month of July 1963, these requirements could be met only on the 2d and on the 14th and in each case for a 10-minute period in the late morning. In addition, the practical requirements of the rocket flight, which yield a profile along the trajectory rather than vertically, dictated the need for stable and uniform ionospheric conditions in the area over which the comparison had to be made. The geometry and timing of this rendezvous experiment is shown in figure 1. Although time simultaneity of the measurements could not be achieved

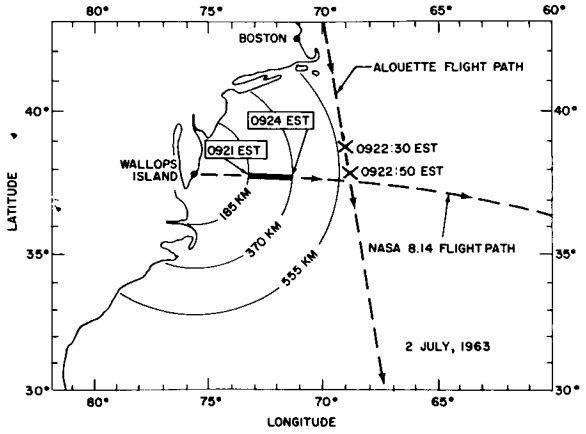


FIGURE 1.—Geometry of the experiment. The heavy portion of the NASA 8.14 flight path indicates the horizontal range over which the ascent rocket data was for altitudes between 300 km and 600 km. Crosses on the Alouette flight path indicate locations and times corresponding to satellite soundings used for the comparison.

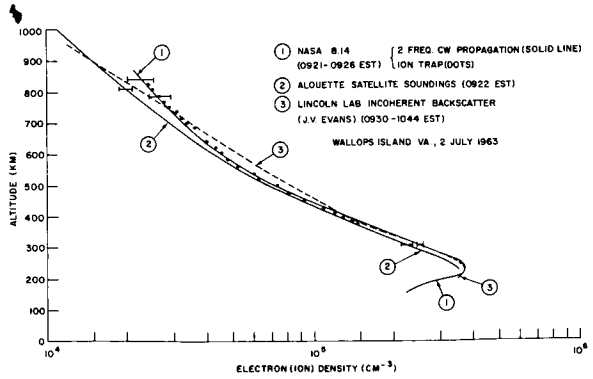


FIGURE 2.—Comparison of charged particle profiles obtained by rocket, satellite and ground-based incoherent-back-scatter measurements. Experimental errors in the rocket and satellite data are discussed in the text.

for the entire rocket flight, it is seen from figure 1 that the ascent data at altitudes between 300 and 600 km (corresponding to the altitude range in which the rocket data are most accurate) were obtained within minutes of the closest satellite data, and at an average horizontal distance of about 300 km. The electron density profiles derived from the Alouette ionograms were virtually identical over the latitude range from 42° N. to 37° N. showing that there were no significant north-south gradients in the topside over a horizontal distance of at least 500 km. In addition preflight monitoring of the ionospheric conditions using bottomside sounders at Wallops Island and at Fort Belvoir had revealed uniformity in the east-west direction. Thus it appears that the requirement for a stable and uniform ionosphere was fairly well met during this test. This factor is also of importance since another comparison was planned with an incoherent backscatter measurement made near Boston, Mass. (42° N.) by J. V. Evans.

The results of this test are given in figure 2, which shows the charged-particle density profiles obtained by Alouette, the rocketborne CW propagation and ion trap experiments and the ground-based incoherent backscatter radar. The rocket data is most accurate for ascent and for altitudes below 700 km, as the result of the trajectory achieved. Above 700 km the horizontal component of the rocket velocity became large compared to the vertical component, which reduced the accuracy of the rocket data. In the case of the CW propagation data an obliquity correction has been made (Jackson and Bauer, 1961) which assumes that the ionosphere is stable and does not have significant horizontal gradients. Assuming an uncertainty of 10% in this obliquity correction, yields an electron density error of 5% at 700 km, increasing to about 15% at 860 km. Below 600 km the estimated error is less than 2%.

The ion trap experiment yields ion densities from the measured collector current by assuming that the rocket velocity is larger than the ion thermal velocity and that the grid transparency for ions in the same as the optical transparency, which is consistent with results of previous flights (Donley, 1963). The maximum error in ion densities is about 5% (due to calibration and telemetry scaling accuracy) below 700 km and about 10% above 700 km. The ion densities shown in figure 2 include a correction which is required when the ion trap is not in the direction of the rocket velocity vector. In the upper portion of the trajectory the angle of attack of the ion trap becomes larger and this angle becomes an important factor in the analysis. The increased error in ion density above 700 km is due to the uncertainty in our knowledge of this angle of

attack. No ion density could be obtained below 367 km due to saturation of the ion trap.

The agreement between the two rocket-borne measurements, is much better at altitudes above 700 km than could be expected on the basis of the stated uncertainties of either measurement. However, this agreement must be fortuitous since comparison with the topside sounder at altitudes above 700 km reveals differences in density of at least 10%, allowing an uncertainty due to analysis alone of about 5% in the topside sounder profile. A more significant comparison with the topside sounder data can be made below 600 km, where the rocket data is believed to be correct within a few percent; in this region the rocket and the satellite profiles agree within 8% which is only slightly greater than the combined uncertainty of the two measurements. Although F2max was poorly defined on the topside ionograms obtained on this occasion, the height of this peak inferred from the Alouette data appears to be too low by at least 20 km, which may be due partly to errors inherent in the true height analysis and partly to a propagation effect (slight ray bending).

A normalized electron density profile (N/N_{max}) derived from incoherent backscatter measurements, assuming that T_e/T_i, the ratio of electron to ion temperature is constant, was supplied by J. V. Evans. This normalized profile gave a height for the F2 peak almost identical to that derived from the rocket data. The backscatter profile shown in fig. 2 was obtained by adjusting the normalized density profile at the F2 peak to agree with the rocket data. There is excellent agreement between the adjusted backscatter data and the rocket data up to an altitude of about 450 km, implying that T_e/T_i may indeed be constant over this range. The disagreement at higher altitudes could be the result of a height dependence in the T_e/T_i ratio; detailed information on this behavior must await the complete spectral analysis of the incoherent backscatter data.

Plotting the rocket data and the satellite data on a geopotential scale shows an essentially constant logarithmic slope from about 275 to 475 geopotential km, indicating an almost constant scale height, $H'=k(T_e+T_i)/m_i g_0$. Assuming that the mean ionic mass m_i is 16 over this height interval, the upper limit of (T_e+T_i) is 2500°± 150° K. Based upon the flux of solar decimeter radiation, the neutral gas temperature and the assumed equal ion temperature at the time of the test were estimated to be between 700° and 800° K. Thus over this altitude region the T_e/T_i ratio would be of the order of 2.5. A comparison of the rocket and Alouette data were made with model profiles assuming a ternary ion mixture in diffusive equilibrium (Bauer, 1962). A differential least square computer analysis was performed varying the following parameters: T_i, T_e/T_i (assumed to be constant with height), and the ion concentration ratios of He^+/O^+ and H^+/O^+ at the reference level of 300 geopotential km. The best fit was obtained for values of T_i between 750° and 800° K and for T_e/T_i ratios between 2.3 and 2.9. The equal concentration levels inferred from this fit correspond to about 560 km for He^+ and O^+, and about 1600 km for H^+ and He^+. It should be understood that these numbers do not represent actual determinations of the transition levels (since T_e/T_i may be altitude dependent) but rather indicate that the experimental profiles are consistent with a ternary ion mixture model. In view of the fact that the experimental data used for the model comparison are for altitudes less than 1000 km, the accuracy of the upper (He^+ to H^+) transition level may be rather poor even if the assumption of a constant T_e/T_i were correct. From the analysis of the retarding potential data obtained during the rocket flight it may be possible to obtain also the variation of T_e with altitude as well as some additional information concerning ion composition. When these data, as well as the spectral data from the incoherent backscatter become available, a more detailed interpretation of the charge density profile in terms of its structure parameters may become possible.

The above discussion illustrates some of the limitations of the topside sounder data and some of the difficulties encountered in determining the basic geophysical parameters from these profiles without additional information. From the standpoint of accuracy however it can be seen that the topside sounder profile compares well with those of other established techniques. In view of its broad geographic coverage, the topside sounder satellite appears to be the most useful tool at the

present time for the synoptic study of the upper ionosphere, provided the limitations of the technique are recognized and proper care is exercized in the analysis and interpretation of its data.

ACKNOWLEDGMENTS

The authors are very grateful to Dr. J. V. Evans of the MIT Lincoln Laboratory for operating the incoherent backscatter radar during the Alouette test and for providing us with his preliminary results for this comparison.

Messrs. G. H. Spaid, R. J. Hagemeyer, G. A. Talley, and W. M. Daniels of the Goddard Space Flight Center were responsible for the design and execution of the CW Propagation Experiment. The personnel of the NASA Wallops Island Station deserve much credit for achieving the very precise timing of the rocket launch which was required for the experiment.

REFERENCES

BAUER, S. J., "On the Structure of the Topside Ionosphere," *J. Atmos. Sciences*, **19,** 273–278, 1962.

BOURDEAU, R. E., WHIPPLE, E. C., DONLEY, J. L., and BAUER, S. J., "Experimental Evidence for the Presence of Helium Ions Based on Explorer VIII Satellite Data," *J. Geophys. Res.*, **67,** pp. 467–475, 1962.

CHAPMAN, J., "Topside Sounding of the Ionosphere," *Advances in the Astronautical Sciences*, Vol. 12, pp. 43–63, 1963.

DONLEY, J. L., "Experimental Evidence for a Low Ion Transition Altitude in the Upper Nighttime Ionosphere," *J. Geophys. Res.*, **68,** 2058–2060, 1963.

JACKSON, J. E., and BAUER, S. J., "Rocket Measurement of a Daytime Electron Density Profile up to 620 km," *J. Geophys. Res.*, **66,** 3055–3057, 1961.

POLARIZATION OF SYNCHROTRON RADIATION AND THE QUASI-TRANSVERSE PROPAGATION REGION

LEO J. BLUMLE

Goddard Space Flight Center

Recent observations of synchrotron radiation from trapped electrons injected by the July 9, 1962, nuclear test have shown that the polarization observed after the radiation had propagated through the ionosphere was slightly elliptical. Ochs et al. (1963) concluded from the power observed in the two propagation modes that the source was slightly elliptical (axial ratio 3.9 to 1). Westfold (1959) and Le Roux (1961) have shown that synchrotron radiation would be composed of two linearly polarized components, one along the magnetic field and one perpendicular to the field. Both authors have considered an isotropic flux over the effective radiation pattern of the electrons, and the results of their work indicate that the total power for a fission spectrum at this frequency would be distributed between the two components, 88 per cent of the total power being in the component perpendicular to the field. Therefore we should observe a much larger power ratio (7 to 1). If the pitch-angle distribution were not isotropic such that there were fewer electrons with higher pitch angles, the power ratio would be even higher.

The apparent discrepancy between the observations and theoretical work (Peterson and Hower, 1963) can be resolved by consideration of the depolarization that results when the propagation conditions are near the quasi-transverse (QT) region (Ratcliffe, 1959).

Under quasi-longitudinal (QL) propagation conditions a linearly polarized wave, upon entering the ionosphere splits into two modes, each of which propagates with different phase velocities. The mode polarizations are approximately circular and therefore, as the wave propagates, the differential mode phase velocity results in the well-known Faraday rotation of the plane of polarization (Daniels and Bauer, 1959; Blumle, 1962). As the QT zone is approached, however, the problem becomes more complex. When we use the accepted magneto-ionic notation (Ratcliffe, 1959) and ignore collisions, the QL approximation for the refractive index is considered valid if

$$Y_T^4/4Y_L^2 \ll 1-X \qquad (1)$$

where

$$Y_L = (f_H \cos\theta)/f$$

$$Y_T = (f_H \sin\theta)/f$$

$$X = f_N^2/f^2$$

f = wave frequency.

f_N = plasma frequency.

f_H = electron gyrofrequency.

θ = angle between the wave normal and the magnetic field.

For the QT region, the sense of the inequality is reversed. Figure 1 is a plot of the ratio

$$\frac{Y_T^4/4Y_L^2}{1-X}$$

versus degrees from transverse for $X=0.04$ and $Y=0.02$, conditions which are typical for 50-Mc/s observations near the magnetic equator. For other conditions, the abscissa is simply proportional to $\gamma^2 \sin^2\theta \tan^2\theta/(1-X)$.

It can be seen from this plot that the QL approximation for the index is valid for propagation

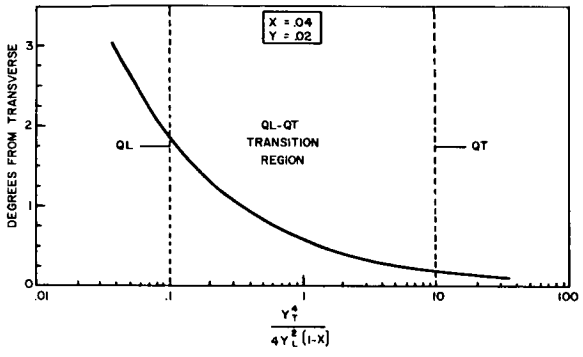


FIGURE 1.—The QL–QT approximations for the phase refractive index.

at angles greater than about 1.82°, where

$$Y_T^4/4Y_L^2(1-X)=0.1 \qquad (2)$$

This conclusion is not entirely correct, however, when we consider the mode polarizations. Figure 2 is a plot of the axial ratio of the polarization ellipse. It can be seen that the QL condition for the refractive index is achieved much closer to the transverse than the condition for circular mode polarization (unity axial ratio) associated with the QL approximation.

In fact, at 1.8° from transverse the axial ratio is only 0.72 and approaches 0.9 only for angles greater than 6° from transverse. Since all the Jicamarca measurements (Ochs et al., 1963) of the synchrotron radiation were made 3° or less from the transverse region, it is generally not possible to deduce the polarization of the source because of this propagation phenomenon. This is possible only if the differential phase path is exactly an integral multiple of the wavelength. Since the antenna used for these observations had a 1° by 2° elliptical beam pattern, energy would also be received in adjacent ray paths, preventing the observation of this special case.

On the basis of this discussion, it appears that the observed polarization of the synchrontron radiation is not in disagreement with the work of Westfold and Le Roux. The apparent discrepancy is due to the depolarization of the radiation as it passes through the ionosphere.

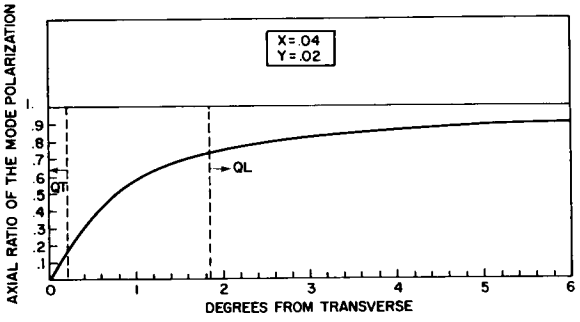


FIGURE 2.—Mode polarizations near the QT propagation region.

The initial polarization cannot be uniquely determined from ground-based observations unless a much narrower beam antenna is used in conjunction with a detailed study of the mode polarizations and differential phase shift. It should be possible to verify this propagation phenomenon by using observations of the ordinary and extraordinary modes from a beacon satellite as the satellite passes through the QT region. To date, no such observations have been made.

ACKNOWLEDGMENTS

I am indebted to Dr. M. P. Nakada and Dr. S. J. Bauer of the Goddard Space Flight Center for their suggestions and discussions about this problem.

REFERENCES

DANIELS, F. B., and BAUER, S. J., The ionospheric Faraday effect and its applications, *J. Franklin Inst.*, **267**(3), 187–200, March 1959.

BLUMLE, L. J., Satellite observations of the equatorial ionosphere, *J. Geophys. Res.*, **67**(12), 4601–4605, November 1962.

LE ROUX, E., Theoretical study of synchrotron radiation of radio sources, *Ann. Geophys.*, **24**, 71, 1961.

OCHS, G. R., FARLEY, D. T., Jr., and BOWLES, K. L., Observations of synchrotron radio noise at the magnetic equator following the high-altitude nuclear explosion of July 9, 1962, *J. Geophys. Res.*, **68**(3), 701–711, February 1, 1963.

PETERSON, A. M., and HOWER, G. L., Synchrotron radiation from high-energy electrons, *J. Geophys. Res.*, **68**(3), 723–734, February 1, 1963.

RATCLIFFE, J. A., *The Magneto-Ionic Theory and Its Applications to the Ionosphere*, Cambridge University Press, 1959.

WESTFOLD, K. C., The polarization of synchrotron radiation, *Astrophys. J.*, **130**, 241, 1959.

ROCKET AND SATELLITE INVESTIGATIONS OF THE IONOSPHERE

ROBERT E. BOURDEAU

Goddard Space Flight Center

GENERAL

In the triennium, a wealth of information was gathered on the temporal, spatial, and energy distribution of low-energy particles in all subdivisions of the ionosphere including the previously relatively unexplored regions above the F_2 peak.

THE D REGION UNDER QUIET SOLAR CONDITIONS

Nicolet and Aikin (1960) proposed that the *D* region (50 to 85 km) is ionized by cosmic radiation below 70 km and in the 70- to 85-km region by Lyman-α radiation acting on a minor constituent, nitric oxide. On the other hand, Poppoff and Whitten [1962] suggest that the upper *D* region is characterized by a lower charged particle concentration resulting entirely from 2 to 8A X-rays. Smith [1961] and Aono et al. (1961) using rocket-borne dc probes with different methods of analyses reported positive ion density values systematically higher than even the Nicolet-Aikin model. Whipple (1960), working with rocket measurements of ion conductivity (Bourdeau et al., 1959), computed a negative ion density profile showing an abundance higher than all theoretical estimates including those of Moler (1960).

Important contributions to *D*-region physics have come from long-term satellite observations which show the Lyman-α flux to be relatively constant with solar activity whereas the flux of 2 to 8A X-rays is extremely variable (Kreplin et al., 1962). Solar radiation data important to *D*-region ionization also are being obtained from the Orbiting Solar Observatory (Lindsay, 1962) and the Ariel satellite (Pounds and Willmore, 1962).

THE DISTURBED D REGION

An empirical correlation has been found between sudden ionospheric disturbances and the times when the energy flux of 2 to 8A X-rays measured on the Greb satellite exceeded a critical value of 2×10^{-3} erg cm^{-2} sec^{-1} (Friedman, 1962). Rocket measurements of electron collision frequency and density were obtained during the early phases of a polar cap absorption (Kane, 1960). A time history of *D*-region electron density profiles during an entire polar cap absorption event has been proposed by Machlum and O'Brien (1962), by inserting proton fluxes measured on the Injun satellite into a conventional equation of ionization and correlating the results in time with riometer absorption observations.

ION CONTENT OF THE LOWER IONOSPHERE

Altitude-dependent models of the production rates of individual ionic species, principally N_2^+, O_2^+, and O^+, have been proposed by Watanabe and Hinteregger (1962) from rocket data obtained by the latter on solar ultraviolet intensities. These are considered qualitative because of the uncertainty in atmospheric composition. Some insight into the important ionospheric reactions can be made by comparing these results with rocket observations of the relative ion abundance. The rocket ion spectrometer results of Johnson et al. (1958) which were obtained at auroral latitudes now have been extended to midlatitudes

by Taylor and Brinton (1961). The latter results confirm the predominance of the diatomic ions (NO^+, O_2^+) below 200 km and of O^+ above this altitude but place a greater emphasis on NO^+.

ELECTRON DENSITIES IN THE LOWER IONOSPHERE

Hinteregger and Watanabe (1962), again using Hinteregger's rocket observations of solar radiation intensity, suggested the relative importance of individual wavelength bands to the formation of the classical ionospheric regions, their most controversial conclusion being a contradiction that E-region ionization is due to soft X-rays. New rocket methods of electron density determination were developed, principally an RF probe operating above the plasma frequency (Jackson and Kane, 1960) and an asymmetric Langmuir probe (Smith, 1961). With the latter device, Smith (1962) has studied the nighttime E region, finding a two-order magnitude decrease from average daytime electron densities. Values of 3 and 1×10^3 electrons cm^{-3} were measured in the 100- to 115-km region at 22 h and near sunrise, respectively. These nighttime profiles show a pronounced valley above 115 km.

Smith (1962) experimentally defined one type of sporadic E in terms of an ionization enhancement of about a factor of 4 confined to a depth of less than 0.5 km with a horizontal dimension greater than 72 km. He suggests that these characteristics are similar to those which led Whitehead (1960) to propose that some types of E_s are caused by redistribution of electrons by wind shear rather than local changes in electron production or loss.

The availability of electron density profiles obtained from rockets flown above the F_2 peak resulted in new attempts at empirical models which interrelate the shape, the maximum electron density, and the "scale height" of the F_2 peak. These empirical models vary in complexity from the use of a constant scale height (Wright, 1960), a scale height gradient (Yonezawa and Takahashi, 1959), and a variable scale height gradient (Chandra, 1963). The controversy over the relative influences of diffusion and radiative recombination near the F_2 peak continues (cf. Sagalyn et al., 1963).

THE UPPER IONOSPHERE

Satellite measurements of charged particle density were made by the use of ion traps (Bourdeau, 1961) and RF probes (Bourdeau and Bauer 1962; Sayers et al., 1962). Sayers et al., working with profiles constructed from Ariel satellite data, claim to have discovered ionization ledges at all latitudes between 0° and 55° varying from 1200 km at local midnight down to 600 km at midday. On the other hand, charged particle density profiles with practically constant electron-ion scale heights between 400 and 700 km were obtained at midlatitudes from rockets by radio-propagation methods (Berning, 1960; Jackson and Bauer, 1961; Bauer and Jackson, 1962; Knecht and Russell, 1962), by the use of ion traps (Hanson and McKibbin, 1961; Hale, 1961), and by RF probes (Ulwick and Pfister, 1962). Their preferred conclusion is that the upper ionosphere is isothermal and in diffusive equilibrium with changes in electron-ion scale height due to corresponding changes in the ionic mixture. Knecht et al. (1961) obtained important diagnostic information on field-alined irregularities from a rocket test of the topside sounding technique in a spread-F condition.

Hanson (1962a), working with Hale's ion density profile, concluded that a 2000-km-thick helium ion layer lies between the upper F region (where O^+ is dominant) and the protonosphere. The first direct evidence for the importance of helium ions was obtained by the use of an ion trap flown on Explorer 8 (Bourdeau et al., 1962). Mange (1960) set forth the governing equations for the electron density distribution in a multiconstituent ionosphere in diffusive equilibrium. Bauer (1962, 1963) proposed that the thickness of the helium ion layer should be strongly dependent on the atmospheric temperature. This has been corroborated by Willmore et al. (1962) who, by the use of an ion trap on the Ariel satellite, report He^+ predominant from 950 km to apogee (1200 km) for a daytime condition, but with protons becoming important near apogee at night. The large variability of upper ionosphere ion transition altitudes is also confirmed by the preliminary results from the Alouette topside sounder satellite (Nelms, 1963; King, 1963; Knecht and Van Zandt, 1963). Bourdeau (1963)

points out that the end result is a higher electron density above 1000 km at night than during the day.

ELECTRON TEMPERATURES

Theortical calculations of electron temperatures based on the assumption of solar radiation as the only ionizing source were made by Hanson and Johnson (1961) and more recently by Hanson (1962b) and Dalgarno et al. (1962). Observational data have been reviewed by Bauer and Bourdeau (1962). Serbu et al. (1961), using a gridded Langmuir probe on Explorer 8, obtained values of 1000°±200°K and 1800°±200°K as the extremes of the diurnal electron temperature variation above 450 km during November 1960. Aono et al. (1961), using a rocket-borne resonance probe, obtained electron temperature values which caused them to conclude temperature equilibrium at E-region altitudes. Spencer et al. (1962) obtained electron temperature data from a symmetric Langmuir probe at altitudes between 100 km and the F_2 peak. They concluded for an auroral and/or a disturbed ionosphere that significant departures from temperature equilibrium could be expected at all altitudes of their flights. The data from one of the flights were obtained at midlatitudes in a quiet ionosphere and show lower electron temperatures in the E and F_2 regions than the other three flights. Willmore et al. (1962) have measured the diurnal electron temperature variation in the isothermal region above 400 km for May 1962, and propose a latitude dependence with midday values of 1200°K at the equator and 1600°K at 55°, all having a pronounced high value during the sunrise period.

ELECTRON CONTENT

Radio transmissions from satellites have been used extensively for determination of electron content between observing sites and altitudes well above the F_2 peak. Yeh and Swenson (1961) reveal strong diurnal and anomalous seasonal variation as well as a depression of electron content during magnetic storms. Garriott (1960) reports gross effects of large magnetic storms on the electron content and the ratio of the electron content above the F_2 peak to the content below. Ross (1960) emphasizes strong seasonal effects and magnetic control of the summer ionosphere. Blumle (1962) obtains a diurnal variation at the magnetic equator which is empirically in phase with the diurnal variation of atmospheric temperature.

SATELLITE RECEPTION OF VERY LOW FREQUENCY

Reception of VLF signals on spacecraft was obtained for the first time. Cain et al. (1961) detected lightning induced whistlers on the Vanguard 3 satellite and emphasized low nighttime absorption by the ionosphere and a high occurrence near the equator. Leiphart et al. (1962) reported reception of VLF on the Lofti satellite from transmitters on the Earth's surface and also noted less attenuation at night with reception as far away as 10,000 miles. They observed little attenuation with altitude and 18-kc time delays from 10 to 200 msec. Rorden et al. (1962) reported Lofti reception of 18-kc signals virtually everywhere by the whistler mode in the ionosphere, with signal attenuation dependent on the direction of the transmitter. They observed one-hop echo delays of 20 to 50 msec, and strong two-hop whistler echoes sometimes exceeding the direct signal at the satellite.

REFERENCES

Aono, Y., Hirao, K., and Miyazaki, S., Positive ion density, electron density and electron temperature in the ionosphere, *J. Radiation Res. Lab., Japan*, **8**, 453–465, 1961.

Bauer, S. J., On the structure of the topside ionosphere, *J. Atmospheric Sci.*, **19**, 276–278, 1962.

———, Helium ion belt in the upper atmosphere, *Nature*, **197**, 1963.

———, and Bourdeau, R. E., Upper atmosphere temperatures derived from charged particle observations, *J. Atmospheric Sci.*, **19**, 218–225, 1962.

———, and Jackson, J. E., Rocket measurement of the electron density distribution in the topside ionosphere, *J. Geophys. Res.*, **67**, 1675–1677, 1962.

Bering, W. W., A sounding rocket measurement of electron densities to 1500 kilometers, *J. Geophys. Res.*, **65**, 2589–2594, 1960.

Blumle, L. J., Satellite observations of the equatorial ionosphere, *J. Geophys. Res.*, **67**, 4601–4605, 1962.

Bourdeau, R. E., Ionospheric results with sounding rockets and satellites, *Space Research II*, edited by H. C. van de Hulst, C. de Jager, and A. F. Moore, pp. 554–573, North-Holland, Amsterdam, 1961.

———, Ionospheric research from space vehicles, *Space Sci. Rev.*, in press, 1963.

———, and BAUER, S. J., Structure of the upper atmosphere deduced from charged particle measurements on rockets and the Explorer 8 satellite, *Space Research III*, edited by W. Priester, North-Holland, Amsterdam, to be published, 1962.

———, DONLEY, J. L., WHIPPLE, E. C., and BAUER, S. J., Experimental evidence for the presence of helium ions based on Explorer 8 satellite data, *J. Geophys. Res.*, **67,** 467–475, 1962.

———, WHIPPLE, E. C., and CLARK, J. F., Analytic and experimental electrical conductivity between the stratosphere and the ionosphere, *J. Geophys. Res.*, **64,** 1363–1370, 1959.

CAIN, J. C., SHAPIRO, I. R., STOLARIK, J. D., and HEPPNER, J. P., A note on whistlers observed above the ionosphere, *J. Geophys. Res.*, **66,** 2677–2680, 1961.

CHANDRA, S., Electron density distribution in the upper *F* region and related atmosphere parameters, *J. Geophys. Res.*, **68,** 1937–1942, 1963.

DALGARNO, A., MCELROY, M. B., and MOFFETT, R. J., Electron temperatures in the ionospheres, *Geophys. Corp. Am. Tech. Rept.* 62-11-N, 1962.

FRIEDMAN, H., Solar radiation, *Astronautics*, **8,** 14–23, 1962.

GARRIOTT, O. K., The determination of ionospheric electron content and distribution from satellite observations, *J. Geophys. Res.*, **65,** 1139, 1960.

HALE, L. C., Ionospheric measurements with a multigrid retarding potential analyzer, *J. Geophys. Res.*, **66,** 1954, 1961.

HANSON, W. B., Upper atmosphere helium ions, *J. Geophys. Res.*, **67,** 183–188, 1962a.

———, Electron temperatures in the upper atmosphere, *Space Research III*, edited by W. Priester, North-Holland, Amsterdam, to be published, 1962b.

———, and JOHNSON, F. S., Electron temperatures in the ionosphere, *Les Congrès et Colloques de l'Université de Liège*, **20,** 390–424, 1961.

———, and MCKIBBIN, D. D., An ion-trap measurement of the ion concentration profile above the F_2 peak, *J. Geophys. Res.*, **66,** 1667–1671, 1961.

HINTEREGGER, H. E., and WATANABE, K., Photoionization rates in the *E* and *F* regions, 2, *J. Geophys. Res.*, **67,** 3373–3392, 1962.

JACKSON, J. E., and BAUER, S. J., Rocket measurement of a daytime electron density profile up to 620 km, *J. Geophys. Res.*, **66,** 3055–3057, 1961.

———, and KANE, J. A., Ionospheric measurements using environmental sampling techniques, *Space Research I*, edited by H. Kallmann-Bijl, pp. 328–339, North-Holland, Amsterdam, 1960.

JOHNSON, C. Y., MEADOWS, E. B., and HOLMES, J. C., Ion composition of the arctic ionosphere, *J. Geophys. Res.*, **63,** 443–444, 1958.

KANE, J. A., Reevaluation of ionospheric electron densities and collision frequencies derived from rocket measurements of refractive index and attenuation, *NASA Tech. Note D-503*, 1960.

KING, J. W., Preliminary studies of the upper ionosphere deduced from topside sounder data, *Nature*, in press, 1963.

KNECHT, R. W., and RUSSELL, S., Pulsed radio soundings of the topside of the ionosphere in the presence of spread *F*, *J. Geophys. Res.*, **67,** 1178–1182, 1962.

———, and VAN ZANDT, T. E., Some early results from the ionospheric topside sounder satellite, *Nature*, in press, 1963.

———, ———, and RUSSELL, S., First pulsed radio soundings of the topside of the ionosphere, *J. Geophys. Res.*, **66,** 3078–3082, 1961.

KREPLIN, R. W., CHUBB, T. A., and FRIEDMAN, H., X-ray and Lyman-alpha emission from the Sun as measured from the NRL SR-1 satellite, *J. Geophys. Res.*, **67,** 2231–2253, 1962.

LEIPHART, J. P., ZEEK, R. W., BEARCE, L. S., and TÓTH, E., Penetration of the ionosphere by very-low-frequency radio signals, *Proc. IRE*, **50,** 16–17, 1962.

LINDSAY, J. C., Solar radiation in the ultraviolet and X-ray spectrum, *Intern. Symp. on Space Phenomena* (IRE/NASA/AEC), 1962.

MAEHLUM, B., and O'BRIEN, B. J., Solar cosmic rays of July 1961, and their ionospheric effects, *J. Geophys. Res.*, **67,** 3281–3288, 1962.

MANGE, P., The distribution of minor ions in electrostatic equilibrium in the high atmosphere, *J. Geophys. Res.*, **65,** 3833–3834, 1960.

MOLER, W. F., VLF propagation effects of a *D*-region layer produced by cosmic rays, *J. Geophys. Res.*, **65,** 1459–1468, 1960.

NELMS, G. L., Scale heights of the upper ionosphere from topside soundings, *Nature*, in press, 1963.

NICOLET, M., and AIKIN, A. C., The formation of the *D*-region of the ionosphere, *J. Geophys. Res.*, **65,** 1469–1483, 1960.

POPPOFF, I. G., and WHITTEN, R. C., *D*-region ionization by solar X-rays, *J. Geophys. Res.*, **67,** 2986–2989, 1962.

POUNDS, K., and WILLMORE, A. P., X-ray measurements on the Ariel satellite, *Proc. Intern. Conf. on the Ionosphere*, in press, 1962.

RORDEN, R., SMITH, R. L., and HELLIWELL, R. A., An interpretation of Lofti VLF observations, *AGARD, Munich*, September 1962.

ROSS, W. J., The determination of ionospheric electron content from satellite Doppler measurements, *J. Geophys. Res.*, **65,** 2601, 1960.

SAGALYN, R. C., SMIDDY, M., and WISNIA, J., Measurement and interpretation of ion density distributions in the daytime *F* region, *J. Geophys. Res.*, **68,** 199–211, 1963.

SAYERS, J., ROTHWELL, P., and WAGER, J. H., Evidence for a further ionospheric ledge above the F_2-region, *Nature*, **195,** 1143–1145, 1962.

SERBU, G. P., BOURDEAU, R. E., and DONLEY, J. L., Electron temperature measurements on the Explorer 8 satellite, *J. Geophys. Res.*, **66,** 4318–4319, 1961.

SMITH, L. G., Electron density measurements by the asymmetric probe, *J. Geophys. Res.*, **66,** 2562, 1961.

———, Rocket measurements of electron density and temperature in the nighttime ionosphere, *J. Geophys. Res.*, **67,** 1658, 1962.

SPENCER, N. W., BRACE, L. H., and CARIGNAN, G. R., Electron temperature evidence for nonthermal equilibrium in the ionosphere, *J. Geophys. Res.*, **67,** 157–175, 1962.

TAYLOR, H. A., and BRINTON, H. C., Atmospheric ion composition measured above Wallops Island, Virginia, *J. Geophys. Res.*, **66,** 2587–2588, 1961.

ULWICK, J. C., and PFISTER, W., Spatial and temporal variations of electron density from an orbiting satellite, *Space Research III*, edited by W. Priester, North-Holland, Amsterdam, to be published, 1962.

WATANABE, K., and HINTEREGGER, H. E., Photoionization rates in the *E* and *F* regions, 1, *J. Geophys. Res.*, **67,** 999–1006, 1962.

WHIPPLE, E. C., Direct measurements of ion density and conductivity in the *D*-region, *XIth Intern. Astron. Congress*, 99–102, 1960.

WHITEHEAD, J. D., Formation of the sporadic-*E* layer in temperate zones, *Nature*, **188,** 567, 1960.

WILLMORE, A. P., BOYD, R. L. F., and BOWEN, J. P., Some preliminary results of the plasma probe experiments on the Ariel satellite, *Proc. Intern. Conf. on the Ionosphere*, in press, 1962.

WRIGHT, J. W., A model of the *F*-region above $h_{max}F_2$, *J. Geophys. Res.*, **65,** 195, 1960.

YEH, K. C., and SWENSON, G. W., Ionospheric electron content and its variation deduced from satellite observations, *J. Geophys. Res.*, **66,** 1061–1068, 1961.

YONEZAWA, T., TAKAHASHI, H., and ARIMA, Y., A theoretical consideration of the electron and ion density distributions in the lower portion of the *F*-region, *J. Radiation Res. Lab., Japan*, **6,** 21–40, 1959.

STRUCTURE OF THE UPPER ATMOSPHERE DEDUCED FROM CHARGED PARTICLE MEASUREMENTS ON ROCKETS AND THE EXPLORER VIII SATELLITE

R. E. BOURDEAU AND S. J. BAUER

Goddard Space Flight Center

The ion composition measured directly at altitudes above the F2 peak by the Explorer VIII satellite (1960 ξ1) is compared with the ion composition inferred indirectly from recent rocket measurements of charged particle densities. These data show that there are two transition regions (from O^+ to He^+ and from He^+ to H^+) in the upper ionosphere rather than a single transition from O^+ to H^+ as was previously believed. The results place the altitude at which $O^+/He^+ = 1$ between 800 and 1400 km depending on the atmospheric temperature. The measured temperature in the upper ionosphere was found to be constant with altitude within a few percent and consistent with a previously developed empirical relation which predicts the temperature as a function of diurnal time and of solar activity. The determined altitudes of the ion transition levels are in good agreement with a theoretical model which predicts these altitudes as a function of atmospheric temperature.

Langmuir probe measurements of electron temperature made on the Explorer VIII satellite, together with those obtained on recent rocket flights, are compared herein with reference atmospheres. This comparison favors the concept of temperature equilibrium in all but the lower F region of the quiet daytime ionosphere. A revision of theoretical considerations, based on present knowledge of ionizing radiation and energy transfer mechanisms, is offered to explain the observed detailed altitude dependence of the difference between electron and neutral gas temperatures. From a comparison of Explorer VIII data and several rocket measurements of charged particle densities, a diurnal temperature variation of about 80 percent is indicated in the isothermal altitude region of the upper atmosphere.

INTRODUCTION

The structure of the upper atmosphere is defined in terms of its density, temperature, and chemical composition. This paper presents results of studies on the structure of the *ionized* atmosphere. The value of the charged particle data is enhanced when, as is done here, the data are compared with recent reference atmospheres and solar radiation observations.

To establish a basis for discussion of the most recent results, it will be helpful to summarize our knowledge of upper atmosphere ionic composition as it existed a year ago. The ions formed in the greatest numbers in the lower ionosphere are N_2^+, O_2^+, and O^+. The N_2 ions disappear very rapidly at low pressures by dissociative recombination,[1] so that their concentration is small. Chemical reactions of O^+ with molecular nitrogen produce NO^+, and the principal ions which exist below the F2 peak are O^+, NO^+, and O_2^+. Early flights of a Bennett RF mass spectrometer in the auroral zone showed that the ions below 200 km are principally diatomic, O_2^+ predominating at the lower altitudes.[2] More recently, flights of the same experiment at middle latitudes have shown that, of the two molecular ions, NO^+ was predominant below 200 km.[3] In both sets of data, the ionic composition measured above 200 km was essentially atomic in nature—mainly O^+. Results from RF spectrometer[4] and ion trap[5] experiments flown on Sputnik III (1958 δ2) showed that O^+ remains

predominant up to at least 800 km. In the early reporting of data[6,7] from a retarding-potential experiment flown on the NASA Explorer VIII satellite (1960 ξ1), it was demonstrated that O^+ predominates at 1000 km in the daytime ionosphere.

In the absence of experimental results, it has been generally believed that at an altitude of approximately 1300 km the ionic composition would change directly from atomic oxygen to protons. One of the important results of the NASA ionospheric physics program has been the conclusion, drawn from several experimental observations, *that an additional transition region must be considered and that a "helium layer" is interposed between the regions where* O^+ *and* H^+ *predominate.* Nicolet[8] had previously deduced from observations of drag on the NASA Echo I satellite (1960 ι1) that neutral helium is an important constituent at very high altitudes. His estimates of the neutral helium number density have subsequently been verified by ground-based optical experiments conducted in the USSR.[9]

Even though the discovery of the ionized helium layer is recent, it is already possible to deduce a preliminary relationship between the atmospheric temperature and the altitudes of the O^+ to He^+ and He^+ to H^+ transition regions. Throughout this paper, we define the transition altitudes as those altitudes where the ratios O^+/He^+ and He^+/H^+ equal unity. Four separate measurements of the lower transition altitude and one of the upper transition altitude are available.

IONIC COMPOSITION

We shall consider first the data from the Explorer VIII retarding-potential experiment, which is illustrated schematically in figure 1. This experiment is based on the proposition that, because of the high satellite velocity, the kinetic energy (relative to the vehicle) of an ion is proportional to its mass. A measurement of this kinetic energy can be derived from the behavior of the collected ion current as a function of an applied retarding potential. Specifically, the potential of the collector, relative to the plasma, at which one half of the ions of mass M_+ are retarded is given by

$$\phi_R = [M_+(V\cos\theta)^2/2e] \qquad (1)$$

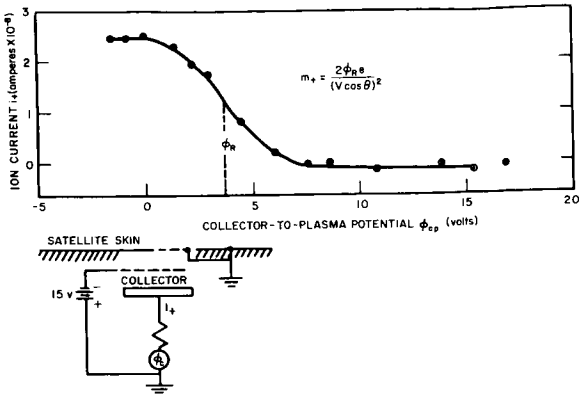


FIGURE 1.—Results of the Explorer VIII retarding-potential experiment at an altitude of 1000 km under daytime conditions. The instrument is shown schematically at the left.

where V is the satellite velocity, θ is the angle of the sensor relative to the velocity vector, and e is the electronic charge. To give accurate ratios of the ionic constituents, the sensor must be pointed in the direction of motion—a condition which, because of the short active life to the satellite, did not prevail except in the altitude region between 700 and 1600 km, and then under daytime conditions only. Experimental points for an altitude of 1000 km are shown in figure 1. The monotonically decreasing nature of the curve is charac-

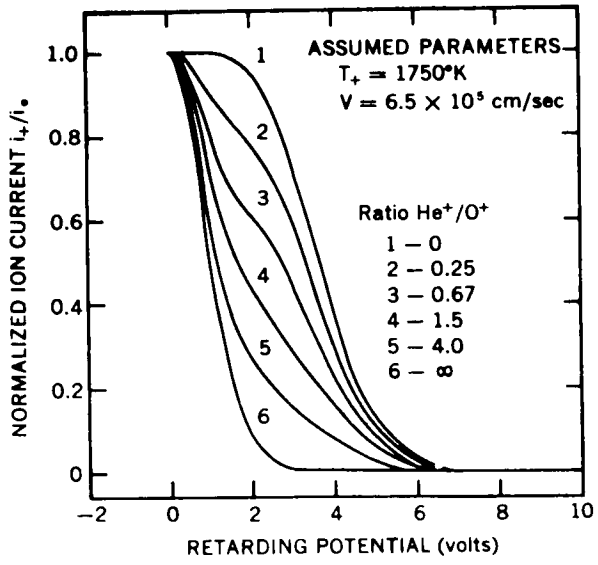


FIGURE 2.—Theoretical retarding-potential curves for a binary mixture of helium and oxygen ions.

teristic of a single ionic constituent which, by substituting into equation 1 the known satellite velocity and orientation and the value of ϕ_R from figure 1, is identified as O^+. The abscissa in figure 1 is the collector-to-plasma potential ϕ_{cp} which is the algebraic sum of the applied collector potential ϕ_c and the satellite-to-plasma potential ϕ_{sp} measured separately by a Langmuir probe.

Theoretical retarding-potential curves computed from an expression given by Whipple[10] for binary mixtures of helium and oxygen and of hydrogen and oxygen are presented in figures 2 and 3, respectively. It is seen that an oxygen-helium mixture is characteristically identified by an inflection point, and an oxygen-hydrogen mixture by distinguishable plateaus. The shapes of these curves are relatively insensitive to the ion temperature T_+. Since the Explorer VIII data at altitudes of about 1600 km are characterized by inflection points, it was concluded that the predominant ions at this altitude are O^+ and He^+.[11] By fitting the experimental points to the family of oxygen-helium curves shown in figure 3, it was found that the lower transition altitude (O^+ to He^+) was about 1400 km for an atmospheric temperature of approximately 1750°K.

Hanson,[12] who first reported on upper atmosphere helium ions, has indirectly determined both transition altitudes from the changes in scale

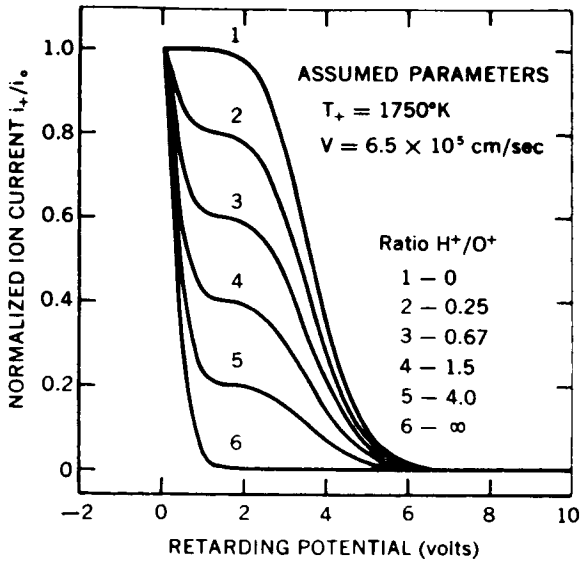


FIGURE 3.—Theoretical retarding-potential curves for a binary mixture of hydrogen and oxygen ions.

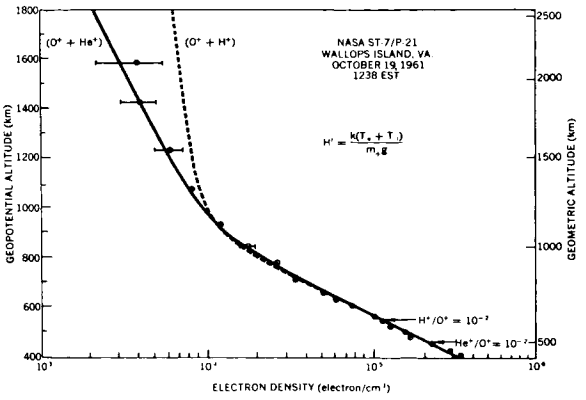


FIGURE 4.—Comparison of the electron density profile obtained by a radio propagation experiment with that of a theoretical model.

height of an ion density profile obtained by Hale[13] in an ion trap experiment flown on NASA Scout ST–2. The atmospheric temperature derived from the scale height of the electron-ion gas in the region between 1600 and 3400 km, on the assumption of a mean ionic mass of 4 amu, was 1600°K. The transition altitudes from O^+ to He^+ and from He^+ to H^+ which Hanson estimated are 1150 and 3500 km, respectively. Figure 4 shows an electron density profile obtained by Bauer and Jackson[14] from a radio propagation experiment flown on NASA Scout ST–7. The right-hand ordinate scale is the true or geometric altitude, while the left-hand scale is the geopotential height which takes into account the altitude variation of the acceleration of gravity. As illustrated, the experimental data are more consistent with a transition from O^+ to He^+ (solid line) than from O^+ to H^+ (dashed line). In this case, the inferred atmospheric temperature is 1350°K and the transition altitude (O^+ to He^+) is 1050 km.

Most recently, Donley (private communication) has made a direct measurement of He^+/O^+ from a retarding-potential experiment flown on NASA Scout ST–9 into a nighttime ionosphere. From a preliminary data analysis, the transition altitude appears to be below 800 km at a time when the atmospheric temperature was approximately 800°K.

Figure 5 illustrates the atmospheric temperature dependence of the transition altitudes as determined from Bauer's theoretical mode[15] for

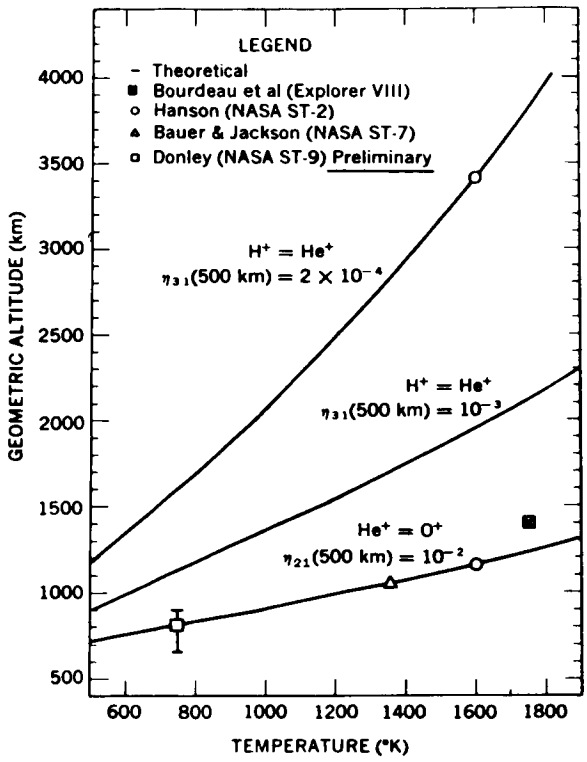


FIGURE 5.—Ion transition altitudes as functions of atmospheric temperature.

the electron density distribution in an isothermal, three-constituent ionosphere in diffusive equilibrium. Three curves are shown: two for the upper transition altitude ($H^+ = He^+$) and one for the lower ($He^+ = O^+$). For the latter case, the prediction assumes that the ratio $He^+/O^+ \equiv \eta_{21}$ has the value 10^{-2} at 500 km, in accordance with the experimental results of Bauer and Jackson.[14] Plotted on the graph are the four experimental results referred to earlier, which show reasonably good agreement with the theoretical curve when we consider that the relative concentrations of these ionic constituents may also vary with temperature at the reference altitude.

The upper theoretical curve assumes that the ratio $H^+/O^+ \equiv \eta_{31}$ has the value 2×10^{-4} at 500 km, a value representative of Hanson's current estimates of proton concentration in the upper ionosphere. The other upper transition altitude curve ($\eta_{31} = 10^{-3}$) is included to illustrate the radical reduction in the thickness of the helium layer which would result if the relative proton concentration were increased by a factor of 5 at the reference altitude.

ALTITUDE COMPARISON OF EXPERIMENTALLY OBTAINED ELECTRON TEMPERATURES WITH REFERENCE ATMOSPHERES

It is of considerable importance to compare electron and neutral gas temperatures, since their relationship depends upon many parameters essential to the quantitative confirmation of present theories regarding the formation of the various ionospheric regions. Because direct and indirect measurements of charged particle temperatures have been made under radically different conditions, and because of the limitations of the kinetic gas temperature models, various investigators have provided conflicting answers to the important question of temperature equilibrium between electrons and heavy constituents. As this paper will show, a consistent pattern can be found in the temperature ratio of charged particles to neutral gas by carefully separating the reported data in altitude and possibly inl atitude, and more importantly by treating conditions of quiet and enhanced solar activity as separate cases.

Before proceeding, it is important to define our use of the term "temperature equilibrium." Actually, because in the ionization process the electrons are liberated with high initial energies, their temperature T_e will be higher than that of the neutral gas T but will approach it asymptotically in time, depending on the efficiency of the efficiency of the energy transfer mechanisms. We shall define temperature equilibrium as existing when the difference between T_e and T is smaller than the estimated uncertainties in reference atmospheres and in experimental methods of measuring charged particle temperatures. We can estimate, perhaps optimistically, that for most cases these uncertainties together are about 10% of the absolute value of the kinetic gas temperature.

From presently available data, an altitude comparison of electron and kinetic gas temperature is best obtained by comparing Langmuir probe measurements of electron temperatures with recent reference atmospheres. Langmuir probes have required considerable development in order to overcome problems associated with the dis-

turbance introduced into the medium by a conducting body—problems so complex that early results undoubtedly contain first-order errors. It was not until 1961 that electron temperatures close to accepted kinetic gas values were first reported for the E region, by Japanese investigators,[16] and for the upper ionosphere, from the NASA Explorer VIII satellite.[17]

To perform a valid comparison of electron and kinetic gas temperatures, it is necessary to select electron temperature data representative of the characteristic reference atmosphere conditions of a quiet sun at middle latitudes. Reported electron temperature results from only two rocket flights and one satellite (Explorer VIII) meet these requirements. These data are compared with the 1961 COSPAR International Reference Atmosphere (CIRA) in figure 6. The rocket Langmuir probe data include the averages obtained by the two probes flown simultaneously by the Japanese, and results obtained by the Michigan group using a bipolar probe flown on NASA Rocket 6.04.[18] All data in the isothermal region, including the reference atmosphere, have been normalized according to Priester's decimeter radiation relation[19] to the solar activity conditions prevailing at the diurnal maximum of March 26, 1961 (the data of both the Japan and Michigan rocket flights). Also included on the graph for future discussion is a measurement of the kinetic gas temperature inferred on the basis of temperature equilibrium from a measured electron density profile.[20]

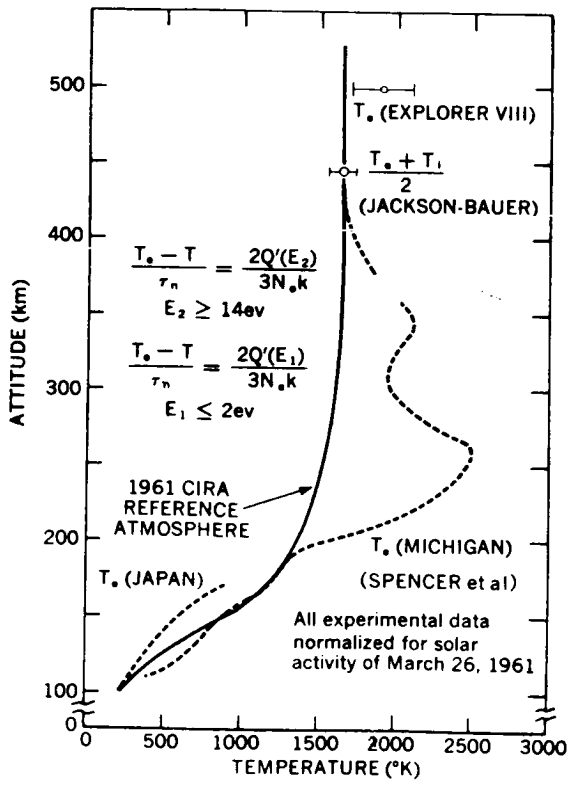


FIGURE 6.—Comparison of the kinetic gas and experimental electron temperatures for a quiet daytime ionosphere at middle latitudes.

In view of the status of Langmuir probe technology and the limitations (imposed by the necessity to assume a neutral composition) of reference atmospheres, the foregoing comparison supports the expectation—based on theoretical considerations of ionizing radiation and energy transfer mechanisms—of temperature equilibrium in all but the lower F region of the quiet daytime ionosphere.[21] The Japanese data actually show lower electron temperatures than the generally accepted kinetic gas temperatures below 170 km. The Michigan values and the reference atmosphere are virtually identical between 140 and 190 km. The Michigan group reports that their values below 140 km have larger uncertainties than their other data. Taken together, then, the two sets of rocket data indicate equilibrium between 100 and about 190 km.

In the F region between 250 and 360 km, the Michigan electron temperature values are sufficiently higher than those of the neutral gas that the difference cannot be ascribed to inadequacies of the reference atmosphere or to experimental electron temperature errors. Consequently, this is a definite indication that departure from temperature equilibrium has been established for the F region, with the maximum electron temperature values occurring at about the altitude of maximum absorption of solar radiation.

The data obtained at apogee of the Michigan flight, which occurred at 360 km (just above the F2 peak), show a trend toward a return to temperature equilibrium. As the next section of this paper will show, it can be predicted by quantitative revisions to the Hanson–Johnson hypothesis that the electron and neutral gas become virtually identical at altitudes between 400 and 500 km. This is indicated by the dashed extra-

polation of the Michigan results in figure 6. There are several experimental justifications for temperature equilibrium well above the F2 peak. Explorer VIII data yield electron temperature values within 15% of those of the neutral gas models. Although this small indicated departure from equilibrium could be real, it is equally likely that it represents inadequacies in the electron temperature measurements. A second justification is the observation from ground-based radar incoherent backscatter experiments[22]—which directly measure the ratio T_e/T_i of electron and ion temperature—that temperature equilibrium prevails near the F2 peak throughout the day except at sunrise and during disturbed ionospheric conditions. The third justification comes from the general agreement of temperatures computed from measured scale-heights of the electron-ion gas above the F2 peak with accepted values of neutral gas temperature in the isothermal region. These data are discussed in more detail in a later section. The value given by Jackson and Bauer[20] is included in figure 6 for comparison with the Langmuir probe data.

Since this comparison is for quiet ionospheric conditions at middle latitudes, it does not include results reported by the Michigan group[18] on three other NASA rocket flights, two obtained under disturbed conditions and one in the auroral region. Also excluded are the results of Smith[23] on a NASA rocket flight made within 24 hours of the onset of a geomagnetic disturbance. The radar incoherent backscatter results[22] have provided experimental evidence that disturbed ionospheric conditions result in values of T_e/T_i of the order of 2.

THEORETICAL CONSIDERATIONS OF THE DIFFERENCE BETWEEN ELECTRON AND KINETIC GAS TEMPERATURES

The most recent quantitative theoretical study of the ionospheric electron temperature and its relationship to the kinetic gas temperature was made by Hanson and Johnson.[21] As figure 7 illustrates, they concluded that the electron and neutral gas temperatures are virtually identical except at altitudes between 160 and 325 km. In this section we shall summarize their hypotheses and then suggest modifications which are more consistent with the data presented in figure 6.

Just after ionization has taken place the newly liberated photoelectrons, which comprise less than 1% of the total electron population, have energies exceeding that of the neutral gas by at least 14 ev. The process by which this excess energy is transferred to upper atmosphere constituents is an altitude-dependent phenomenon which may be described as follows:

Below 225 km, inelastic collisions with neutral particles reduce the photoelectron energy to 2 ev, the cutoff point of the excitation cross-section of atomic oxygen. The 2 ev electrons then share their energy with the ambient electrons, thus raising the temperature of the latter above that of the heavy constituents. Because this process is very fast, the electrons have a Maxwellian energy distribution—a conclusion verified experimentally by the shapes of the volt-ampere curves of those Langmuir probes whose potentials are permitted to reach the plasma potential.

After a Maxwellian distribution of electron energy is established, the temperature difference is calculated by relating the heat input to the electrons to the heat lost by elastic collisions with

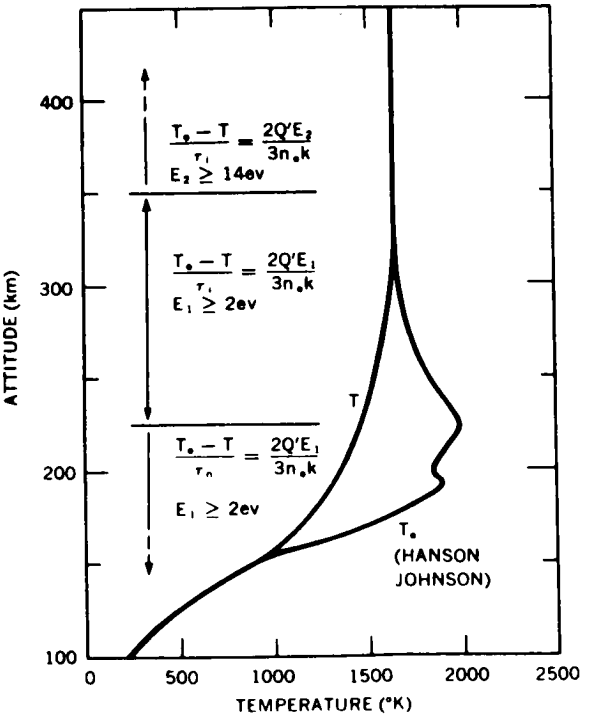


FIGURE 7.—Theoretical considerations of temperature equilibrium for quiet daytime conditions.

heavy particles:

$$2Q'(E_1)/3N_ek=(T_e-T)/\tau_n, \qquad z<225 \text{ km}, \quad (2)$$

where Q' is the rate at which photoelectrons of energy E_1(2 ev or less for this case) are released, N_e is the electron density, k is Boltzmann's constant, and τ_n is the time needed for electrons of energy E_1 to transfer their excess energy to neutral particles.

Between 225 *and* 350 *km*, the process is the same except that the time τ_i needed for 2 ev electrons to transfer their excess energies to ions is shorter than τ_n; thus,

$$2Q'(E_1)/2N_ek=(T_e-T)/\tau_i, \qquad 225 \text{ km}<z<350 \text{ km}. \quad (3)$$

Above 350 *km*, the inelastic collision process is no longer efficient, and the photoelectrons transfer their energy directly to the ambient electrons, raising the value of E_1 to E_2, that is, to 14 ev or more depending upon whether two or one photoelectrons are released per incoming photon:

$$2Q'(E_2)/3N_ek=(T_e-T)/\tau_i, \qquad z>350 \text{ km}. \quad (4)$$

Hanson and Johnson calculated from available atmospheric models that temperature equilibrium as we have defined it would prevail except between 160 and 325 km, a region where high solar radiation absorption is accompanied by moderate values for the respective equipartition times. The principal uncertainties in their computations result from corresponding uncertainties in cross sections and densities of the atmospheric constituents. They noted, in proof of this, that an overestimated excitation cross section of atomic oxygen caused them to overestimate the altitude at which inelastic collisions become ineffective; and we shall note below that this radically affects the altitude domains in which the various energy transfer mechanisms come into play. The rearrangement offers one explanation for the experimental results presented in figure 6.

The major effect of lowering the altitude above which inelastic collisions are unimportant is that the efficiency of energy transfer by elastic collisions with ions is greatly reduced. If we must consider equipartition times τ_i based on electrons with energies of 14 ev or greater rather than 2 ev at all altitudes, we estimate that energy transfer to ions does not control the electron temperature below about 600 km.

Following this reasoning, we must now consider two altitude domains below 600 km: an upper portion where energetic electrons of 14 ev or greater transfer their energy directly to the ambient electrons; and a lower portion where, because of the intervening inelastic collision process, there are only 2 ev of energy available for selective electron heating. In both domains the temperature difference (T_e-T) is finally controlled by elastic collisions with neutral constituents.

In order to provide new estimates of (T_e-T) at all altitudes, $Q'(E)$ was calculated with (T_e-T) values from figure 6 in the altitude region where the difference is measurable (200–360 km); equilibrium times τ_n in accordance with Hanson and Johnson; and electron densities measured during the Michigan flight by an ionosonde and by a rocket-borne propagation experiment. In these calculations, a discontinuity appears in the $Q'(E)$ function at 325 km. This discontinuity may be attributed to the transition altitude where inelastic collisions are no longer efficient: Above that altitude more energy (E_2) is available for selective electron heating; and the secondary maximum in the Michigan electron temperature profile (fig. 6) in the 300–350 km region may possibly be explained in this fashion. The new equations which seem to apply below and above 325 km are included as part of figure 6.

By an extrapolation of the $Q'(E)$ function, and of the τ_n values given by Hanson and Johnson (now *reduced* above 325 km), and of the electron density profile, we conclude for the ionospheric conditions represented by figure 6 that the electron and kinetic gas temperatures are virtually identical below about 190 km and above about 450 km. Below 190 km, the justification is the reduction of photoelectron energy by inelastic collisions together with high collision frequencies. In the higher altitude region, it appears that the heat input to the electrons is decreasing more rapidly with altitude than is the combined effect of an increasing equipartition time and a decreasing electron density. This does require a somewhat more rapid decrease in $Q'(E)$ at the higher altitudes than would be inferred from a recent study by Watanabe and Hinteregger[24]; but, as

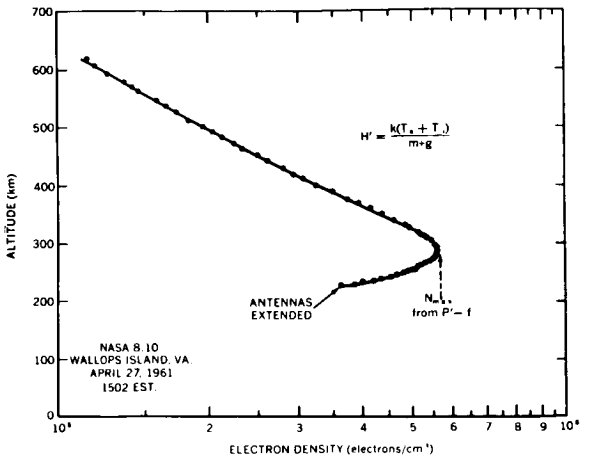


FIGURE 8.—Electron density profile from a radio-propagation experiment, illustrating the isothermality of the upper ionosphere.

they point out, their analysis is a first approximation which can be refined as the atmospheric composition and some photoionization and absorption cross-sections become better known.

DIURNAL AND SOLAR ACTIVITY VARIATION OF UPPER IONOSPHERE TEMPERATURES

Neutral gas temperatures above 200 km are generally derived from an assumed atmospheric composition together with atmospheric densities computed from satellite drag observations. The drag observations show that density variations are correlated with solar activity. Although not the source of upper atmosphere heating, solar decimeter radiation—which is observable at the earth's surface—is an indicator of this interrelationship. Different empirical equations relating the 10.7 and 20 cm solar radiation atmospheric temperature in the isothermal altitude region have been derived by Jacchia[25] and Priester.[19] Jacchia's equations are based on an atmospheric model by Nicolet[26] which includes the presence of helium and in which the mean molecular weight is computed on the basis of diffusive equilibrium of the atmospheric constituents. Priester's model, on the other hand, makes use of a molecular mass variation typical of the 1961 CIRA reference atmosphere.

Theories of upper atmospheric heating can be improved by comparing such models of the diurnal and solar activity variations of neutral gas temperatures with charged particle temperatures obtained from measurements in the isothermal altitude region. To do so, it is necessary to assume temperature equilibrium well above the F2 peak; this assumption was justified theoretically and experimentally in the previous sections.

One method of deducing charged particle temperatures above the F2 peak is to measure accurately the electron or ion density profile. From theoretical considerations as well as experimental evidence, it is now well established that the distribution of electrons and ions at these altitudes generally corresponds to a diffusive equilibrium distribution. One such experimental evidence, a day-time electron profile measured by a radio propagation method,[20] is illustrated in figure 8. In such cases the slope of the charged particle distribution is a unique measure of the scale height of the electron-ion gas; and for regions where one ionic constituent predominates it is also a measure of the sum of the electron (T_e) and ion (T_i) temperatures. In general, the scale-height of the electron-ion gas is given by

$$H' = k(T_e + T_i)/m_+ g = [d/dz\ (\ln N)]^{-1}, \quad (5)$$

where m_+ is the mean ionic mass, g is the acceleration of gravity, N is the charged particle density, and z the altitude. Thus, H' is a measure of $(T_e + T_i)$ or, in the case of temperature equilibrium, of the neutral gas temperature $T = (T_e + T_i)/2$.

Parenthetically, the high degree of isothermality in $(T_e + T_i)$ evidenced by the experimental results in figure 8 provides additional support for the assumption of temperature equilibrium well above the F2 peak. For this isothermality to occur with radical differences between T_e and T_i requires the rather unlikely condition that the energy input to the electrons $Q'(E)$ decreases with altitude identically as the combined rate of increase in equipartition time and of decrease in electron density.

Six rocket measurements of the altitude profile of charged particle densities above the F2 peak have been reported during the last year. In addition to those presented in figures 3 and 7, two electron density profiles were obtained from the

TABLE 1.—*Rocket Measurements of Atmospheric Temperature above the F2 Peak*

Rocket flight	Local time (hours)	Temperature (°K)	Source of data
NASA ST–7	1238	1350±100	Bauer and Jackson (ref. 14)
NASA ST–2	1500	1600±100	Hanson (ref. 12).
NASA 8.10	1502	1640±90	Jackson and Bauer (ref. 20).
NASA 8.15	1812	1200±60	Jackson et al. (ref. 27).
NASA 8.04	2044	1240±70	Hanson and McKibbin (ref. 28).
NASA 8.17	2326	800±40	Jackson et al. (ref. 27).

NASA topside sounder program,[27] and the remaining two are ion density profiles. The six sets of data are listed in table 1 along with the kinetic gas temperature inferred from an assumption of temperature equilibrium.

In figure 9 are plotted the Jacchia (dashed lines) and Priester (solid lines) models of the atmospheric temperature in the isothermal altitude regions as functions of solar decimeter radiation flux. For comparison, the temperatures from table 1 are located on the graph according to the solar decimeter radiation observed on the individual launch dates. Also included is a direct measurement of the neutral gas temperature obtained by Blamont et al.[29] from a sodium release experiment flown on NASA Rocket 8.05. It is important to note that the rocket data have not been normalized for diurnal time.

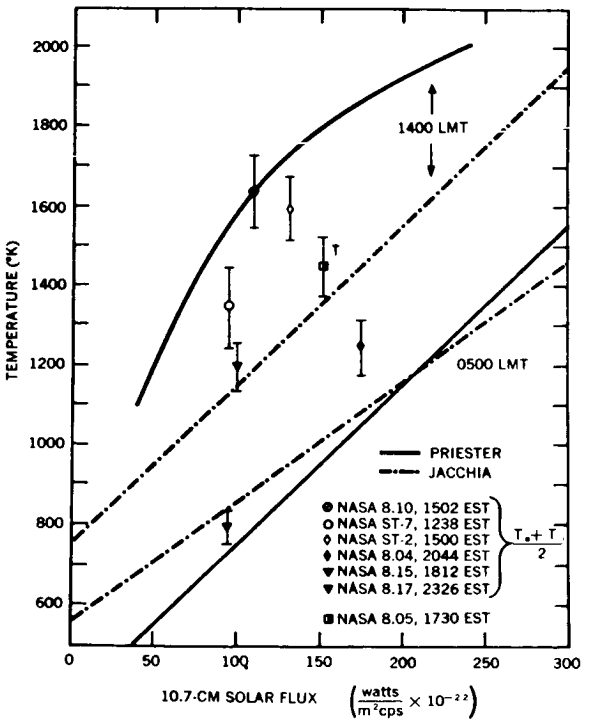


FIGURE 9.—Upper ionosphere temperature as a function of solar activity.

The value of Blamont et al., 1475±40°K, which was obtained at local sunset, is higher than Jacchia's diurnal maximum value, suggesting that the latter's daytime temperatures are too low. One value of $(T_e+T_i)/2$ obtained by a topside sounder experiment (NASA Rocket 8.15) is in very good agreement with the value given by Blamont et al. in that it also was taken at local sunset and it too is higher than Jacchia's diurnal maximum curve. Three of the rocket measurements of $(T_e+T_i)/2$ were taken within 2 hours of the diurnal maximum. All three are consistent with the data of Blamont et al. in that they are higher than the temperatures in Jacchia's model but lower than or equal to Priester's values for the diurnal maximum. The remaining measurements of $(T_e+T_i)/2$ are for nighttime conditions. One taken close to midnight (NASA Rocket 8.17) agrees quite well with both Jacchia's and Priester's values for diurnal minimum. This should be expected since little cooling will occur between midnight and the diurnal minimum. The other nighttime value (NASA Rocket 8.04) was taken about 2 hours after sunset when the atmosphere was cooling toward diurnal minimum. As expected, this value falls between the diurnal maximum and minimum values.

Considering that both Jacchia's and Priester's

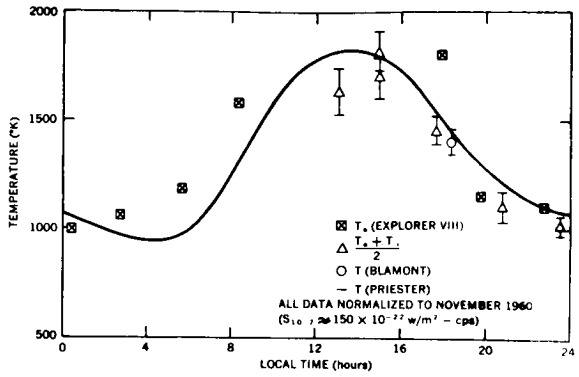


FIGURE 10.—Diurnal variation of upper ionosphere temperatures.

kinetic gas temperatures are inferred rather than measured, the general agreement of the various charged particle measurements with their temperature models is reasonably good. It appears from the daytime values of temperatures deduced from charged particle observations, as well as from the direct measurement of neutral gas temperature given by Blamont et al., that the diurnal variation of temperature in the isothermal region is closer to the 80% suggested by Priester than to the 35% suggested by Jacchia.

Inasmuch as the rocket temperature measurements are in somewhat closer agreement with Priester's model than with Jacchia's, they are plotted in figure 10 as a function of local mean time to illustrate further the agreement with his implied diurnal variation.[30] Electron temperatures reported from the Explorer VIII satellite are also included. All of these data have been normalized to the average 10.7-cm flux for the month of November 1960. It should be emphasized that the individual electron temperature values from Explorer VIII have a 200°K error spread which is not indicated on the graph. Assuming temperature equilibrium, it would appear from the Explorer VIII data that Priester's daytime values are too low. However, there is a possibility of second-order errors in these Langmuir probe measurements of electron temperature.[31] There is fair agreement between indirect measurements of charged particle temperatures and Priester's values, with a possible implication that the diurnal maximum is broader than indicated.

VARIABILITY EXTREMES OF ELECTRON DENSITIES IN THE UPPER IONOSPHERE

It has been brought out indirectly in the preceding section that electron densities in the upper ionosphere are controlled by the atmospheric temperature and ion composition. They are, of course, additionally controlled by N_{max} values at the F2 peak, which are in turn governed by absorption of solar radiation and by recombination processes in the lower F region. In this section, to illustrate the variability extremes which these factors produce in upper ionosphere electron densities, we shall compare ionosonde data with electron density measurements by J. A. Kane (private communication) using an RF probe experiment flown on Explorer VIII.

With ionosonde data for N_{max} values and with the then prevailing diurnal variation of atmospheric density, figure 11 shows the extreme theoretical electron density profiles which—assuming diffusive equilibrium and temperature equilibrium—would be expected during Explorer VIII's active lifetime. The shaded areas indicate the variability of electron densities observed by the RF impedance probe. Each segment plotted at the nearest 100 km level represents about 50 data points.

In the RF probe experiment, which was originally developed by Jackson and Kane,[32] the inflight capacitance C of a shortened dipole antenna is compared to the latter's free space capacitance C_0 at a radio frequency f. The electron density is computed from the simplified Appleton–Hartree formula which relates N_e to the dielectric

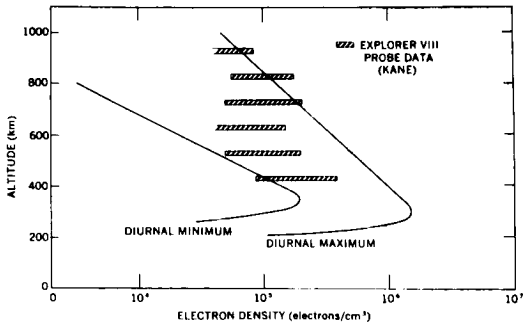


FIGURE 11.—Comparison of electron densities from the Explorer VIII RF impedance probe with theoretical models.

constant K of the medium as follows:

$$K = C/C_0 = 1 - (81N_e/f_2), \qquad (6)$$

where f is in kilocycles.

The principal uncertainty in this measurement is due to the ion sheath which forms about the antenna; the error can be estimated from a knowledge of the potential that the spacecraft acquires relative to the medium.[33] In the case of Explorer VIII, where satellite potentials varied between approximately zero for daytime conditions and -1 volt at night, Kane estimates that the uncertainties due to the ion sheath corresponds to electron densities of the order of 2×10^4 electrons/cm^3. For this reason, values below 4×10^4 electrons/cm^3 were not considered in the data recorded in figure 11.

It was observed that whenever the satellite, *at perigee* (425 km), was within 1 degree of latitude and longitude of an ionosonde, the electron density observed on the satellite was consistent with what would be expected from the N_{max} value. On some perigee transits occurring near local midnight it was noted that fluctuations in the electron density were less than 10% over a distance of the order of 500 km along the direction of the satellite orbit.

As would be expected, the experimental values at the lower altitudes, which correspond to a nighttime condition, are in close agreement with the theoretical diurnal minimum curve. The values which exceed those of the model above 700 km are attributed to the fact that in this region the satellite was passing either into a sunrise or sunset condition; in the former case diffusive and probably temperature equilibrium may not apply.

ACKNOWLEDGMENTS

The authors are indebted to J. L. Donley for the use of his retarding potential data and to J. A. Kane for the use of his RF probe data.

REFERENCES

1. Faire, A. C., and Champion, K. S. W., "Measurements of Dissociative Recombination and Diffusion in Nitrogen at Low Pressures," *Phys. Rev.*, **113**(1): 1–6, January 1, 1959.
2. Johnson, C. Y., Meadows, E. B., and Holmes, J. C., "Ion Composition of the Arctic Ionosphere," *J. Geophys. Res.*, **63**(2): 443–444, June 1958.
3. Taylor, H. A., Jr., and Brinton, H. C., "Atmospheric Ion Composition Measured above Wallops Island, Virginia," *J. Geophys. Res.*, **66**(8): 2587–2588, August 1961.
4. Istomin, V. G., "An Investigation of the Ionic Composition of the Earth's Atmosphere Using Rockets and Satellites," in: *Artificial Earth Satellites*, ed. by L. V. Kurnosova, New York: Plenum Press, 1960, Vol. 2, pp. 40–44.
5. Krassovsky, V. I., "Exploration of the Upper Atmosphere with the Help of the Third Soviet Sputnik," *Proc. IRE*, **41**(2): 289–296, February 1959.
6. Bourdeau, R. E., "Ionospheric Results with Sounding Rockets and the Explorer VIII Satellite," in: *Space Research II: Proc. 2d Internat. Space Sci. Sympos., Florence, April* 1961; ed. by H. C. van de Hulst, C. de Jager, and A. F. Moore, Amsterdam: North-Holland Publ. Co., 1961, pp. 554–573.
7. ———, Donley, J. L., and Whipple, E. C., Jr., "Instrumentation of the Ionosphere Direct Measurements Satellite (Explorer VIII)" NASA Technical Note D-414, April 1962.
8. Nicolet, M., "Helium, an Important Constituent in the Lower Exosphere," *J. Geophys. Res.*, **66**(7): 2263–2264, July 1961.
9. Shefov, N. N., "Émission de l'Hélium dans la haute Atmosphère," *Annales de Géophysique*, **17**(4): 395–402, 1961.
10. Whipple, E. C., Jr., "The Ion-Trap Results in 'Exploration of the Upper Atmosphere with the Help of the Third Soviet Sputnik'," *Proc. IRE*, **47**(11): 2023–2024, November 1959.
11. Bourdeau, R. E., Whipple, E. C., Jr., et al., "Experimental Evidence for the Presence of Helium Ions Based on Explorer VIII Satellite Data," *J. Geophys. Res.*, **67**(2): 467–475, February 1962.
12. Hanson, W. B., "Upper-Atmosphere Helium Ions," *J. Geophys. Res.*, **67**(1): 183–188, January 1962.
13. Hale, L. C., "Ionospheric Measurements with a Multigrid Potential Analyzer," *J. Geophys. Res.*, **66**(5): 1554, May 1961 (abstract).
14. Bauer, S. J., and Jackson, J. E., "Rocket Measurement of the Electron-Density Distribution in the Topside Ionosphere," *J. Geophys. Res.*, **67**(4): 1675–1677, April 1962.
15. ———, "On the Structure of the Topside Ionosphere," *J. Atmos. Sci.*, **19**(3): 276–278, May 1962.
16. Aono, Y., Hirao, K., and Miyazaki, S., "Positive Ion Density, Electron Density and Electron Temperature in the Ionosphere by the Kappa-8–5 and -6 Rockets," *J. Radio Res. Labs.* (Japan), **8**(40): 453–465, November 1961.
17. Serbu, G. P., Bourdeau, R. E., and Donley, J. L., "Electron Temperature Measurements on the Explorer VIII Satellite," *J. Geophys. Res.*, **66**(12): 4313–4315, December 1961.

18. SPENCER, N. W., BRACE, L. H., and CARIGNAN, G. R., "Electron Temperature Evidence for Nonthermal Equilibrium in the Ionosphere," *J. Geophys. Res.*, **67**(1): 157–175, January 1962.
19. PRIESTER, W., "Solar Activity Effect and Diurnal Variation in the Upper Atmosphere," *J. Geophys. Res.*, **66**(12): 4143–4148, December 1961.
20. JACKSON, J. E., and BAUER, S. J., "Rocket Measurement of a Daytime Electron-Density Profile up to 620 Kilometers," *J. Geophys. Res.*, **66**(9): 3055–3057, September 1961.
21. HANSON, W. B., and JOHNSON, F. S., "Electron Temperatures in the Ionosphere," *Mémoires de la Société Royale des Sciences de Liège*, **4**, 390–424, 1961.
22. BOWLES, K. L., OCHS, G. R., and GREEN, J. L., "On the Absolute Intensity of Incoherent Scatter Echoes From the Ionosphere," *J. Res. NBS*, **66**D(4): 395–407, July–August 1962.
23. SMITH, L. G., "Electron Density Measurements by the Asymmetrical Bipolar Probe," *J. Geophys. Res.*, **66**(8): 2562, August 1961 (abstract).
24. WATANABE, K., and HINTEREGGER, H. E., "Photoionization Rates in the E and F Regions," *J. Geophys. Res.*, **67**(3): 999–1006, March 1962.
25. JACCHIA, L. G., "A Working Model for the Upper Atmosphere," *Nature*, **192**(4808) 1147–1148, December 23, 1961.
26. NICOLET, M., "Density of the Heterosphere Related to Temperature," Smithsonian Institution, Astrophys. Observ., Res. in Space Sci., Spec. Rept. No. 75, September 19, 1961.
27. JACKSON, J. E., KNECHT, R. W., and RUSSELL, S., "First Results in NASA Topside Sounder Satellite Program," Paper presented at 8th Annual Meeting of the American Astronautical Society, Washington, January 1962 (to be published).
28. HANSON, W. B., and MCKIBBIN, D. D., "An Ion-Trap Measurement of the Ion Concentration Profile Above the F2 Peak," *J. Geophys. Res.*, **66**(6): 1667–1671, June 1961.
29. BLAMONT, J. E., LORY, M. L., and SCHNEULER, J. P., "Mesure de la Température de la haute Atmosphere a l'Altitude de 370 km," in: *Space Research II: Proc. 2d Internat. Space Sci. Sympos., Florence, April* 1961, ed. by H. C. van de Hulst, C. de Jager, and A. F. Moore, Amsterdam: North-Holland Publ. Co., 1961, pp. 974–980.
30. JASTROW, R., "Results of Experiments in Space," *J. Aerospace Sci.*, **29**(4): 377–388, 409, April 1962.
31. BOURDEAU, R. E., and DONLEY, J. L., et al., "Measurements of Sheath Currents and Equilibrium Potential on the Explorer VIII Satellite," *J. Astronautical Sci.*, **8**(3): 65–73, fall, 1961.
32. JACKSON, J. E., and KANE, J. A., "Measurement of Ionospheric Electron Densities Using an RF Probe Technique, *J. Geophys. Res.*, **64**(8): 1074–1075, August 1959.
33. KANE, J. A., JACKSON, J. E., and WHALE, H. A., "The Simultaneous Measurement of Ionospheric Electron Densities by CW Propagation and RF Impedance Probe Techniques," NASA Technical Note D-1098, January 1962.

EXPLORER VIII SATELLITE MEASUREMENTS IN THE UPPER IONOSPHERE

R. E. BOURDEAU AND J. L. DONLEY

Goddard Space Flight Center

This is a more extensive report than earlier ones on upper ionospheric ion composition (Bourdeau et al., 1962) and electron temperature (Serbu et al., 1961) and on spacecraft-plasma interaction (Bourdeau et al., 1961), all measured by use of Explorer VIII.

Results from an ion-retarding-potential experiment show that the upper ionospheric composition responds to the neutral gas temperature. Specifically, during the satellite's active life (November–December, 1960), O^+ ions predominated from perigee (425 km) up to about 800 km at night and 1500 km at diurnal maximum; the base of the helium ion region was located at 800 km during the sunrise period and at 1500 km in the daytime; the base of the protonosphere possibly was located at 1200 km during the sunrise period and above 1800 km at diurnal maximum.

For the latitudes indicated, the electron temperature (T_e) data—restricted herein to magnetically quiet days ($A_p<15$)—are consonant with current models of the diurnal electron density behavior and with a hypothesis of solar ultraviolet radiation as the only ionizing agent. For the 6-hour period centered at midnight, the average T_e observed at altitudes 425–600 km was 900°K with a standard deviation of 150°K. There were no anomalously high values and no significant change with magnetic dip (0°–75°N). For the period 10–16 hours LMT, at magnetic dips between 50–70°S and at altitudes between 1000–2400 km, the observed average T_e was 1600°K with a standard deviation of 200°K when a small percentage of anomalously high values are excluded. The most pronounced feature of the measured diurnal variation is that T_e exceeds the neutral gas temperature by a factor of about 2.5 during the sunrise period in the altitude region 600–900 km. Explanations are suggested for differences between the Explorer VIII T_e observations and other ground-based and spaceflight measurements of upper ionospheric parameters.

The measured satellite-plasma interaction is consistent with theory, lending confidence to the geophysical results described above. The observed "average" satellite potential varied from a few tenths of a volt negative at night, to zero when the measured daytime charged particle density was 10^4 cm^{-3} and thence to a few tenths of a volt positive for daytime densities of 10^3 cm^{-3}. Superimposed on the "average" potential were experimentally observed potential gradients across the satellite skin—an effect produced by the movement of a conducting body through a magnetic field. The measured orientation sensitivity of three types of current flowing between the satellite and the ionosphere is described.

INTRODUCTION

The Explorer VIII Satellite was devoted primarily to the measurement of upper ionospheric parameters by environmental sampling techniques. Prior to this satellite launching, the ionosphere had been studied mainly by classical ground-based and rocket-borne radio-propagation methods. The success of environmental sampling methods depends on evaluation of the interactions between a spacecraft and the surrounding ionized medium. The relative sparsity of spaceflight observations on the interaction phenomena dictated (1) that the spacecraft configuration be as simple as possible and (2) that it carry supporting experiments designed specifically to study the interaction and thus permit valid interpretation of the geophysical data.

The interaction was minimized by restricting the use of protuberances, by foregoing the use of solar cells and so maintaining an equipotential surface, and by limiting the telemetry transmitter

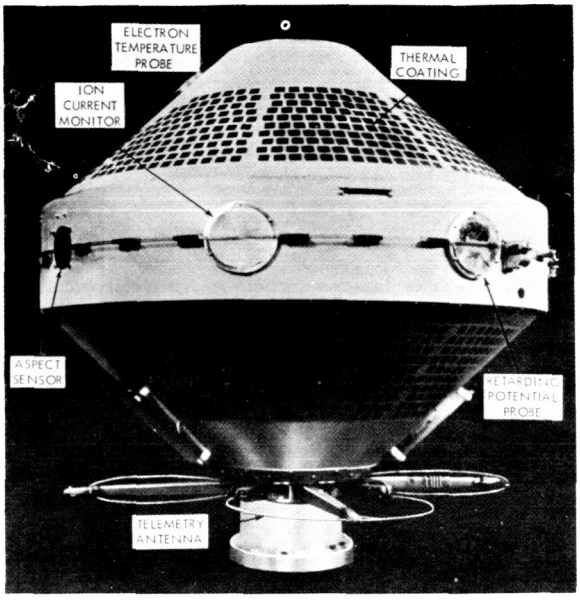


FIGURE 1.—The Explorer VIII satellite.

power to a value which previous results showed would not seriously affect the vehicle potential. These factors proved advantageous in that the studied interaction is quite explainable from kinetic theory considerations as applied to satellites (Gringauz and Zelikman, 1957), thus lending confidence to the reported ionospheric parameters. However, the limited active life together with the nature of real time telemetry transmissions and the scarcity of receiving sites able to accept the low signal levels restricted the acquired data to specific latitude, altitude, and temporal conditions.

The satellite was launched on November 3, 1960, into an orbit with a 50° equatorial inclination, a perigee of 425 km, and an apogee of 2400 km. The planned active life was 2 months. The spacecraft (fig. 1) consisted of two truncated cones joined at the equator. Thermal coatings were placed on both cones in a pattern designed to maintain an equipotential surface. The spin rate at injection was reduced to an orbital value of 22 rpm so that retarding potential curves could be obtained for a minimum orientation change. Of the 10 experiments, 6 are pertinent to this report.

Two different but gratifyingly redundant electron temperature sensors were located near the forward end of the spin axis. One of these also provided data on the ambient electron density, the average potential ϕ_0 of the satellite, and the behavior of the electron current i_e from medium to satellite as a function of position relative to the velocity vector. This electron temperature sensor is visible in figure 1; the other was diametrically opposite. The remaining four sensors were on the satellite's equator; two are seen in figure 1 and the other two are diametrically opposite. One of these was a retarding potential analyzer designed to measure ion composition. The second measured the positive ion current i_+ from medium to satellite as a function of position relative to the velocity vector. The third measured the sum of i_e and the photocurrent i_p. A comparison of the output of this sensor, when not oriented toward the sun, with the i_e measurement made near the forward end of the spin axis reveals the behavior of i_e as a function of magnetic field orientation. The fourth experiment provided redundancy by measuring the sum of i_e, i_+ and i_p, or the total current exchange between the spacecraft and the ionosphere.

THE POTENTIAL OF THE EXPLORER VIII SATELLITE

The "average" potential of a conducting body at rest where RF and magnetic fields and solar radiation may be neglected is given by

$$\phi_0 = -(kT_e/e)\ln(J_e/J_+) \qquad (1)$$

where k is Boltzmann's constant, e is the electronic charge, and J_e and J_+ are the respective electron and ion current densities in the undisturbed ionosphere. On the assumption of temperature and charge equilibrium between electrons and ions ($T_e = T_+$, $n_e = n_+$), this reduces to:

$$\phi_0 = -(kT_e/e)\ln[(m_+)^{1/2}/(m_e)^{1/2}] \qquad (2)$$

where m_+ and m_e are the respective ion and electron masses. In a medium containing only O^+ ions at a temperature of 1000°K, the computed potential is −0.44 volt.

In the practical case of a conducting body moving with satellite velocity through the nighttime ionosphere, the average potential (again neglecting the effects of RF and magnetic fields)

is given by:

$$\phi_0 = -(kT_e/e)\ln\left(\int J_e\, dS_e \Big/ \int J_+\, dS_+\right) \quad (3)$$

where S_e and S_+ are the areas over which the respective electron and ion currents are effective. It can be shown that this reduces approximately to

$$\phi_0 \approx -(kT_e/e)\ln(v_e/V), \quad (4)$$

where v_e and V are the respective electron and satellite velocities. For $T_e = 1000°\mathrm{K}$, the computed average potential is approximately -0.29 volt. This is to be compared with measured Explorer VIII nighttime average potential values between -0.5 and -0.75 volt for the electron temperature indicated. This is an improvement by an order of magnitude in the agreement between measured and predicted satellite potentials over that obtained by other investigators, specifically Krassovsky (1959) who reported negative potentials up to 6 volts from Sputnik III. Factors which might contribute to the difference between predicted and observed Explorer VIII nighttime potentials are the presence of RF and magnetic fields—which are not taken into account in equation 4, and thus affect the predicted result—and a contact potential which can influence the experimental value. The rectification effect of the antennas used for telemetry transmissions would tend to make the vehicle potential more negative than that computed from equation 4. The Japanese work (Aono et al., 1962) on resonance probes demonstrates that if the frequency of RF transmissions exceeds the plasma frequency the effect should be small. Whale (1963), on the other hand, points out that this is so only for low values of RF power. It is possible, therefore, from the work of other investigators to show qualitatively that a large increase in potential would not be expected for the amount of power (100 mw) and the frequency (108 Mc used by the Explorer VIII telemetry system. However, these other data are not quantitative enough to permit a categorical statement that an increase of a few tenths of a volt—which is the difference between the predicted and observed Explorer VIII values—is not possible from the RF rectification effect. The earth's magnetic field can affect the electron diffusion current preferentially. However, for the Explorer VIII form factor and orbit considered here, E. C. Whipple, Jr. (private communication) computes a negligible effect of the magnetic field on the average satellite potential. Another possibility is an unknown in the measurement of the spacecraft potential. This measurement, described in detail in a later section, depends on the assumption that the aperture electrode of the measuring instrument is at the same potential as the adjacent satellite skin. Possible differing work functions could result in different surface contact potentials and consequently introduce an error in the measurement.

Consider now the more complicated case of daytime satellite potentials. Solar radiation causes electrons to be emitted from the satellite surface in such a way that

$$\phi_0 = -(kT_e/e)\ln\left(\int J_e\, dS_e \Big/ \int J_+\, ds_+ + \int J_p\, dS_p\right). \quad (5)$$

For the altitudes considered here, the photocurrent effect $\int J_p\, dS_p$ masks the ion diffusion effect $\int J_+\, dS_+$ and thus competes with electron diffusion $\int J_e\, dS_e$ for the determination of the polarity of the satellite potential. For an orientation where the sun shines directly on the satellite's equator, S_e is approximately $4 \times S_p$. The observed magnitude from Explorer VIII of J_p is 5×10^{-9} amps cm^{-2} (Bourdeau et al., 1961), a value independent of altitude for the particular orbital elements. This is to be compared with a value of 2.3×10^{-9} amps cm^{-2} previously measured by Hinteregger et al., (1959) on a rocket at a lower altitude. The transition from a negative to a positive satellite potential for the Explorer VIII case can be predicted to occur when $4J_e = en_e v_e = 5 \times 10^{-9}$ amps cm^{-2}. For $T_e = 1000°\mathrm{K}$, it is predicted that the polarity of the satellite potential will reverse for 10^3 cm$^{-3} < n_e < 10^4$ cm^{-3}. This is consistent with our observations that the average daytime Explorer VIII potential was approximately -0.15 volts for $n_e \approx 10^4$ cm^{-3} and a few tenths of a volt positive at apogee when $n_e \approx 10^3$ cm^{-3}.

The motion of the satellite with velocity $\vec{V}$ through the magnetic field $\vec{B}$ produces an induced

potential that is a function of position on the satellite surface, where $\vec{d}$ is the vector distance of

$$\phi = \phi_0 + (\vec{V} \times \vec{B}) \cdot \vec{d} \qquad (6)$$

any point on the surface from the satellite center. A satellite potential of ϕ_0 will be measured at all points that lie on a plane through the satellite center perpendicular to $\vec{V} \times \vec{B}$. All other points will be more positive or negative than ϕ_0 as they are situated on one side of this plane or the other (fig. 2). The measurement (Bourdeau et al., 1961) at an altitude of 1000 km shows a potential

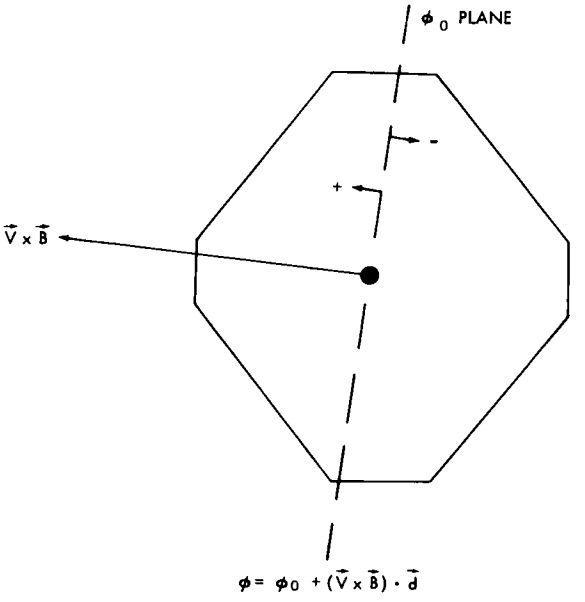


FIGURE 2.—The induced emf effect.

difference of 0.14 volt across the satellite equator, a value consistent with that computed from equation 6 using known values of $\vec{V}$, $\vec{B}$ and $\vec{d}$.

MEASUREMENTS OF PLASMA-TO-SATELLITE ION CURRENT

The positive ion current i_+ flowing from the ionosphere was monitored by the sensor shown schematically in figure 3. The inner grid is biased negatively to suppress photoemission from the collector and to remove incoming electron current so that the collector responds only to ions diffusing from the ionosphere.

As the satellite spins, it is possible to plot i_+ as a function of the azimuth angle of the sensor

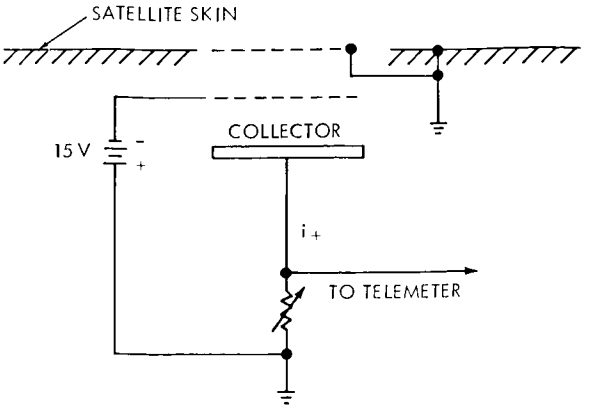


FIGURE 3.—Ion current monitor.

relative to the velocity and solar vectors. A typical example is illustrated in figure 4. The absence of a current when the sensor is pointed at the sun is proof that photoemission from the collector has been successfully suppressed. The observation that i_+ is zero in the satellite wake is to be expected from the relative satellite and ion velocities and is definite but indirect experimental evidence for an electron sheath immediately adjoining the vehicle at this location.

When the sensor is pointed within 45° of the velocity vector, the observed collector current is given by the equation

$$i_+ = \alpha_+ n_+ A e V \cos\theta, \qquad \theta < 45°, \qquad (7)$$

where α_+ is the combined electrical transparency of the grids for positive ions, A is the area of the

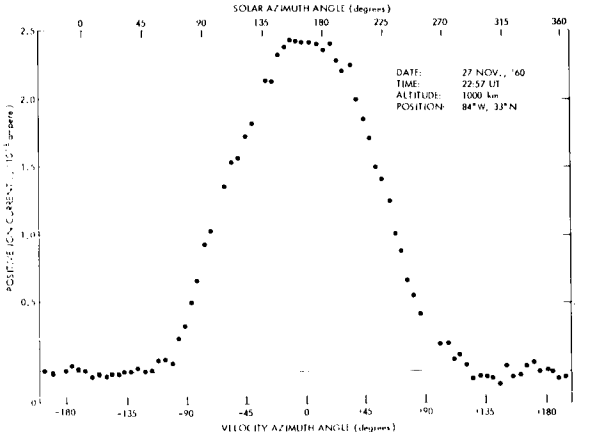


FIGURE 4.—Ion current as a function of aspect.

collector and θ the angle between the sensor normal and the satellite velocity vector. A comparison of the measured i_+ from this sensor with that observed on an exposed collector described in a succeeding section shows that the electrical transparency α_+ is equal to the optical transparency (92%). Whenever, as in this case, the average satellite potential is negative and V is large compared to the thermal velocity of the ion, the ion collection volume is computed from known parameters and consequently the device provides an accurate measurement of charged particle density. Confidence in this statement is based on the agreement between n_+ values computed from equation 7 with measured i_+ values and electron densities measured by the two-element electron temperature experiment described in a succeeding section. An even better test has been performed by Donley (1963), who shows good agreement between ion densities obtained by a similar experiment on a SCOUT rocket and electron densities simultaneously measured by Bauer and Jackson using radio-propagation methods.

When the angle of the sensor normal relative to the velocity vector is 90°, the observed i_+ should behave as though the body is at rest. Current theory for this condition results in

$$i_+ = \alpha_+ n_+ A e a / 2\sqrt{\pi}, \qquad \theta = 90°, \qquad (8)$$

where a is the most probable thermal velocity of the particle. We find that the observed value for $\theta = 90°$ is larger than what would be computed from equation 8 for reasonable values of ion temperature. This discrepancy represents a gap in current theories which assume a discrete boundary for the satellite's surrounding ion sheath. The most likely explanation for the difference is that equation 8 should be modified to account for the effect on the positive ion collection of electric fields penetrating the sheath at the sides of the satellite (Schulz and Brown, 1955). Another possible reason is that factually we are not dealing with the planar geometry for which equation 8 applies.

MEASUREMENTS OF PLASMA-TO-SATELLITE ELECTRON CURRENT

It can be expected that the electron current diffusing from the ionosphere to satellite would be modulated in accordance with orientation changes relative to both the velocity and magnetic field vectors. It is possible to present measurements of these two effects separately by choosing data obtained at critical satellite orientations. Specifically, we can examine measured electron currents taken at small distances from the ϕ_0 plane (cf. Figure 2) where changes due to the magnetic field effect are not permitted even though the spacecraft is spinning. Data were so obtained by use of the experiment illustrated in Figure 5. The sensor consists of two electrodes, an aperture grid maintained at spacecraft potential, and a collector. The positive bias on the latter serves to exclude the effects of positive ion and of photocurrents, resulting in unambiguous measurements of only the electron diffusion current.

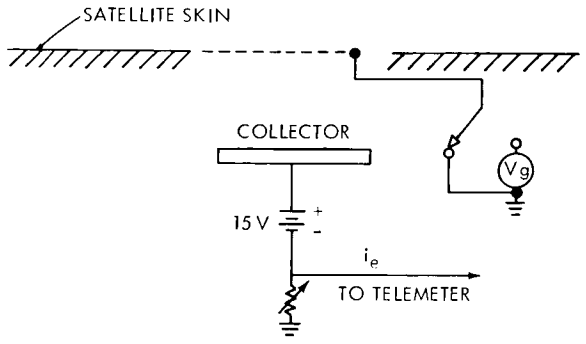


FIGURE 5.—Electron temperature probe.

The result, in the form of a graph of electron current as a function of orientation relative to the velocity vector, taken when the average satellite potential was −0.15 volt, is illustrated in Figure 6. A maximum is observed in the direction of motion and a minimum in the satellite's wake, representing 15 percent more or less, respectively,

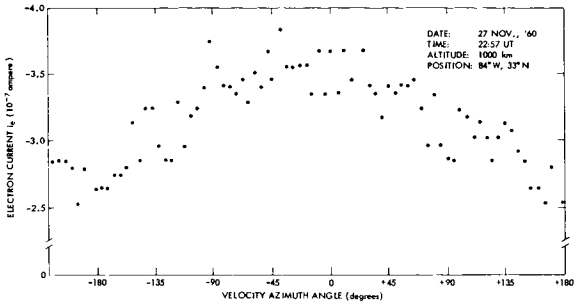


FIGURE 6.—Electron current as a function of aspect.

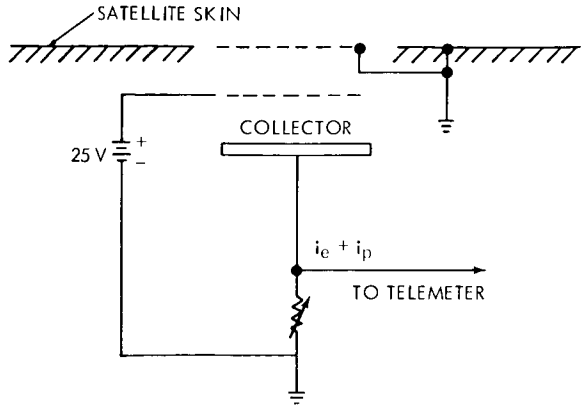


FIGURE 7.—Electron and photocurrent monitor.

than that measured for an orientation perpendicular to the velocity vector. This modulation index is explainable to the first order by the ratio of electron-to-satellite velocity. The results do contrast with those of Willmore et al. (1962) who report at least an order of magnitude electron current depletion in the wake of the Ariel satellite.

We have considered the possibility that the discrepancy between the aforestated Explorer VIII and Ariel observations is due to the fact that the data illustrated in figure 6 were obtained near the forward end of the spin axis and thus represent results taken near the edge of the wake. This consideration involves examining data taken when the spin axis was perpendicular to the velocity vector and obtained by use of an electron current monitor situated on the satellite's equator. A schematic of the electron current monitor is illustrated in figure 7. This sensor responds to photocurrent when pointed toward the sun and

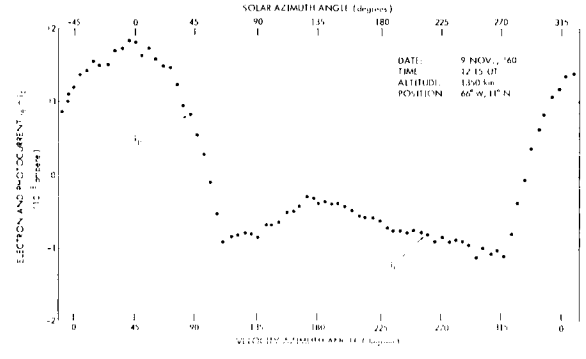


FIGURE 8.—Electron and photocurrent as a function of aspect.

to the diffusion of ambient electrons from the ionosphere. The result illustrated in figure 8 was obtained for an orientation such that magnetic field modulation of the electron diffusion current can be considered minimal. A large photocurrent effect is observed whenever the sensor is oriented within 45° of the sun. Elsewhere it is assumed that we are observing only electron diffusion current. As is expected, this current is a minimum when the sensor is located in the wake. We additionally emphasize (1) that the small percentage modulation is consistent with the relative electron and satellite velocities and, (2) that an appreciable

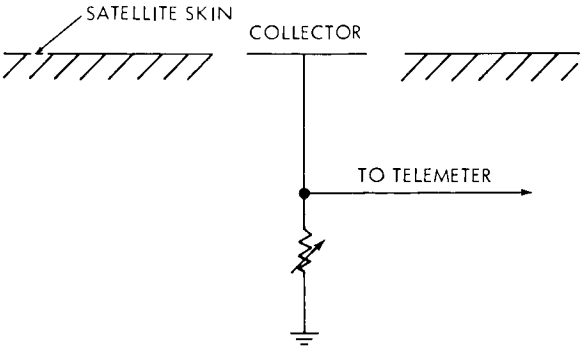


FIGURE 9.—Total current monitor.

current is observed for $\theta = 180°$, when the sensor is extremely close to the *center* of the wake. We still are left with a difference between the Explorer VIII and Ariel results which perhaps is due to the different spacecraft for m factors.

We now can consider magnetic field modulation of the electron diffusion current by examining measured values of total or net current to the satellite surface. This was accomplished by measuring the current to a collector flush with and insulated from the satellite skin (fig. 9). This particular sensor was located on the satellite's equator. A typical daytime result is shown in figure 10. Three pronounced effects are evident: (1) the peak ram ion current as the sensor most nearly points in the direction of motion; (2) the masking photocurrent as the sensor points to the sun; and (3) the maximum electron current

when $\theta \approx 280°$. It is the latter observation which illustrates magnetic field modulation of the electron current. A previous analysis (Bourdeau et al., 1961) shows that this maximum occurs exactly when the azimuth angle between the sensor normal and $\vec{V} \times \vec{B}$ is zero degrees, a position where the satellite surface is driven most positive by the induced emf. Since the electron current is related exponentially to the surface potential, the sharpness of the current variation near this specific orientation is to be expected.

ELIMINATION OF INTERACTION EFFECTS IN DERIVING GEOPHYSICAL PARAMETERS

It is possible from measurements of current flowing to the satellite to extract information on the ambient electron density and temperature and on the ambient ion density and composition. However, the complex behavior of the satellite-plasma interaction, especially short-term variations of satellite potential and unwanted currents require extreme care in so doing. Our methods of deriving geophysical parameters can be described by continuing reference to figure 10 and the discussion which follows.

To extract the ambient ion concentration, we need to know unambiguously the behavior of i_+. It is seen from figure 10 that if we were to use an exposed electrode to determine n_+, we have the problem of accounting for unwanted masking electron and photocurrents. This is most easily accomplished by experimental separation in the manner illustrated schematically in figure 3, leading to the easily interpretable result shown in figure 4. Accurate positive ion densities then are obtainable by use of equation 7 provided that one precisely knows the angle of attack and provided that ϕ_0 is negative. The latter requirement arises from the fact that a positive satellite potential inhibits positive ions from entering the sensor.

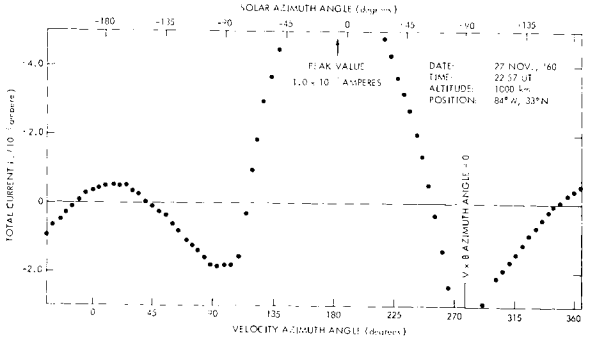


FIGURE 10.—Total current as a function of aspect.

The determination of electron concentration and temperature and of ion composition depends upon a measurement of the electron or ion diffusion current as a function of an applied retarding potential. To accomplish this we must insure that the measured current changes reflect only those due to applied retarding potential. The measurement of ion composition described in the succeeding section was made at the satellite equator. Again, separation of unwanted electron and photocurrents was experimentally accomplished. We next need to account for orientation changes of i_+. This is accomplished by restricting the reduced data to only that obtained when i_+ is orientation invariant, specifically obtaining a volt-ampere curve when the sensor always was closely pointed into the direction of motion.

Similarly, in extracting the electron parameters we first experimentally separate out the unwanted ion and photocurrents. We feel that this is especially important for charged particle densities less than 10^4 cm^{-3}, since from figure 10 we know that the unwanted photocurrent can exceed the electron diffusion current even when the satellite is at plasma potential. Once the unwanted currents are separated out, we only need account for orientation changes in the wanted electron diffusion current. With specific reference to figure 5, it is seen that the electron temperature sensor has two modes of operation. In the first mode, no retarding potential is applied so that the orientation sensitivity illustrated by figure 6 can be measured. From this knowledge, we then can take that volt-ampere curve in the second mode of operation where the electron diffusion current changes entirely reflect that due to the applied potential.

SUMMARY OF ION COMPOSITION RESULTS

Ion composition was measured on the Explorer VIII satellite by use of a sensor located on the equator and identical to that illustrated in figure 3 except that a retarding potential was applied in series with the collector. Previous publications reported that O^+ was the predominant ion at

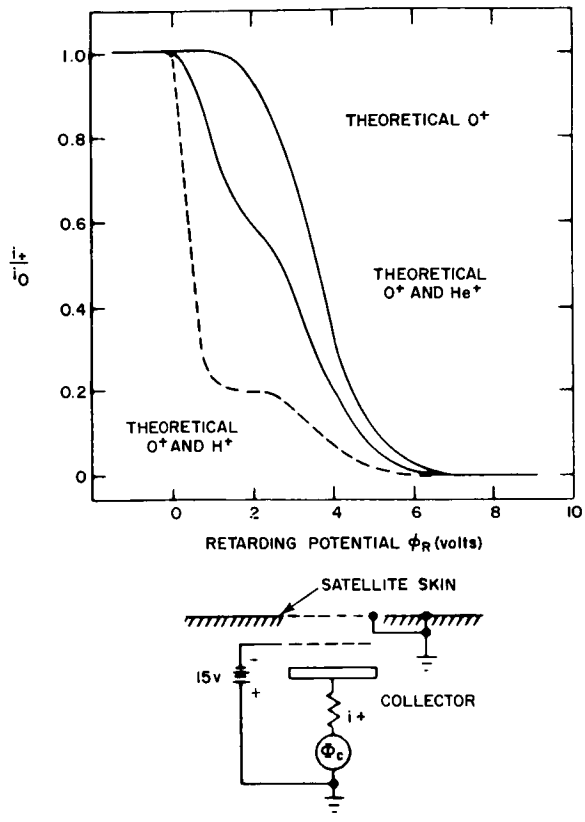


FIGURE 11.—Ion retarding potential experiment.

altitudes below 1000 km in the daytime ionosphere (Bourdeau, 1961) and that helium ions become important above this altitude again for the daytime ionosphere (Bourdeau et al., 1962). In this section, we extend the reported results to infer the diurnal upper ionospheric ion composition behavior for the period November–December 1960.

The principle of the experiment (fig. 11) is based on the fact that an ion's kinetic energy (relative to the satellite) is proportional to its mass. In the upper ionosphere, we are concerned principally with three types of ions: O^+, He^+ and H^+. Whipple (1959) first set forth the theoretical equations for the volt-ampere behavior of planar ion traps. As illustrated in figure 11, the shape of this volt-ampere curve is characterized by an inflection point for a binary mixture of O^+ and He^+ and by a plateau for a mixture of O^+ and H^+. In the illustration, the abscissa is the ion retarding potential, ϕ_r, which is the algebraic sum of the satellite and collector (ϕ_c) potentials. The ordinate is the collector current normalized to that observed at zero retarding potential.

The diurnal behavior of upper ionospheric composition for the months of November and December 1960, is illustrated in figure 12. The two upper volt-ampere curves were taken for different altitudes at times which, from the standpoint of the neutral gas temperature represent diurnal minimum conditions. We observe by comparison that (1) the transition altitude where O^+ and He^+ ions had equal concentrations was at about 770 km, and (2) at 980 km the predominant ion was He^+. In the latter volt-ampere curve, there is slight evidence—from the sharp drop at zero retarding potential—that protons become a trace constituent at 1000 km, so that the transition altitude from He^+ to H^+ would occur at about 1200 km. The lower two curves represent data obtained midway between diurnal minimum and diurnal maximum. We observe from the first that O^+ ions dominated at 800 km and from the second that the O^+–He^+ transition altitude has risen to 1500 km. We have observed some diurnal maximum data where this transition took place as high as 1800 km. Our daytime volt-ampere curves show no trace of protons.

In summary then, the Explorer VIII retarding potential experiment defines an upper ionosphere where, in late 1960, O^+ ions predominated up to

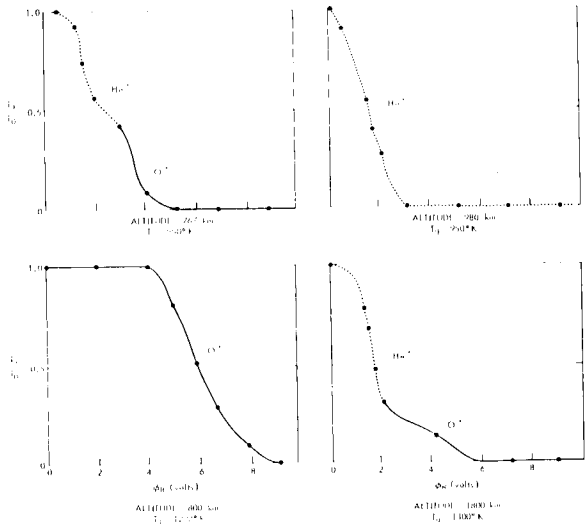


FIGURE 12.—Ion composition results.

about 800 km at night and up to at least 1500 km during midday. He^+ ions predominated from 800 km to at least 1200 km at night and from 1500 to at least 1800 km at midday. This qualitatively agrees with the theoretical work of Bauer (1963) who postulates that these transition altitudes and thus the thickness of the helium ion region can be related directly to the neutral gas temperature T_g. If we accept a value for T_g of 950°K at the diurnal minimum, then the nighttime Explorer VIII composition results correspond quantitatively to Bauer's theoretical expectation. On the other hand, if we accept current reference atmospheres where the diurnal maximum T_g is given at about 1400°K, then the measured O^+–He^+ transition at midday is somewhat higher than in Bauer's model. There is closer agreement with Bauer's model if the daytime electron temperatures discussed in the succeeding section are representative of T_g. The daytime O^+–He^+ transition level is somewhat higher than that inferred by Hanson (1962a) from an ion density profile taken at a similar epoch of the solar cycle.

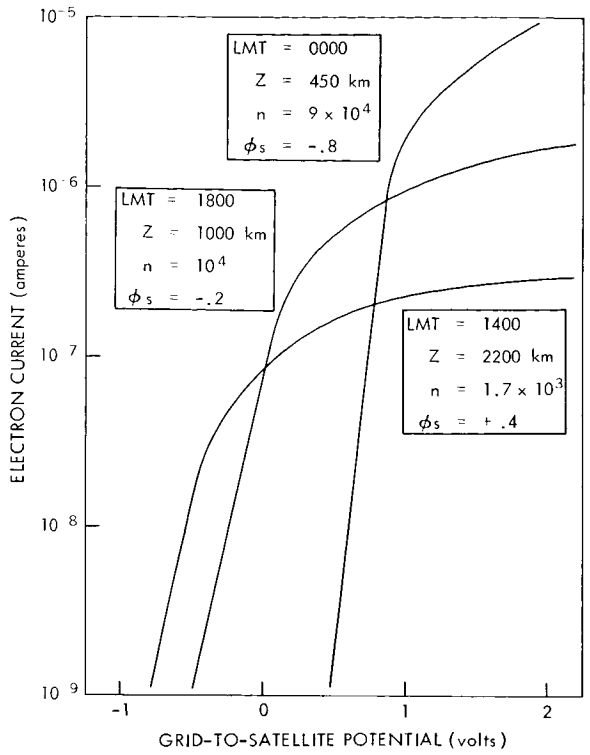


FIGURE 13.—Typical experimental electron temperature volt-ampere curves.

SUMMARY OF ELECTRON TEMPERATURE RESULTS

The Explorer VIII satellite contained two electron temperature probes. The first was the two-electrode device illustrated in figure 5. In the first of this sensor's two modes of operation, the aperture grid was maintained at spacecraft potential and the device was used to monitor the orientation sensitivity of electron diffusion current, a typical example of which is illustrated in figure 6. In the second mode, a retarding potential applied to the aperture grid permitted a measurement of satellite potential and of the ambient electron density and temperature. Typical volt-ampere curves for three conditions are illustrated in figure 13. Electron temperatures are computed from the slope of the exponential portion of the volt-ampere curves. Satellite potential is obtained from the position along the abscissa of the point at which the current departs from its exponential behavior. Electron density is computed from the electron current value taken at this specific retarding potential. The first curve is representative of midnight quiet ionosphere conditions in the 400–600 km region, where values of 900°K, 10^5 cm^{-3}, and -0.75 volt represent typical values of electron temperature and density and satellite potential. The second curve is typical of midday quiet ionosphere conditions at 1000 km where T_e, n_e, and ϕ_s were typically 1600° K, 10^4 cm^{-3} and -0.15 volt. The third curve is typical of midday quiet ionosphere apogee (2400 km) conditions where T_e, n_e, and ϕ_s were typically 1600°K, 10^3 cm^{-3} and $+0.25$ volt.

Although by the method described in previous sections, we have taken into account possible errors in electron temperature determination due to orientation changes of satellite potential and of the three types of exchange currents, there is left a possibility of error due to changes in the aperture grid's electrical transparency with applied potential. This possible error we believe to be small from a comparison of the two-electrode results with a different electron temperature experiment also included on the satellite. Electrically this latter experiment, located near the forward end of the spin axis, was identical to that illustrated in

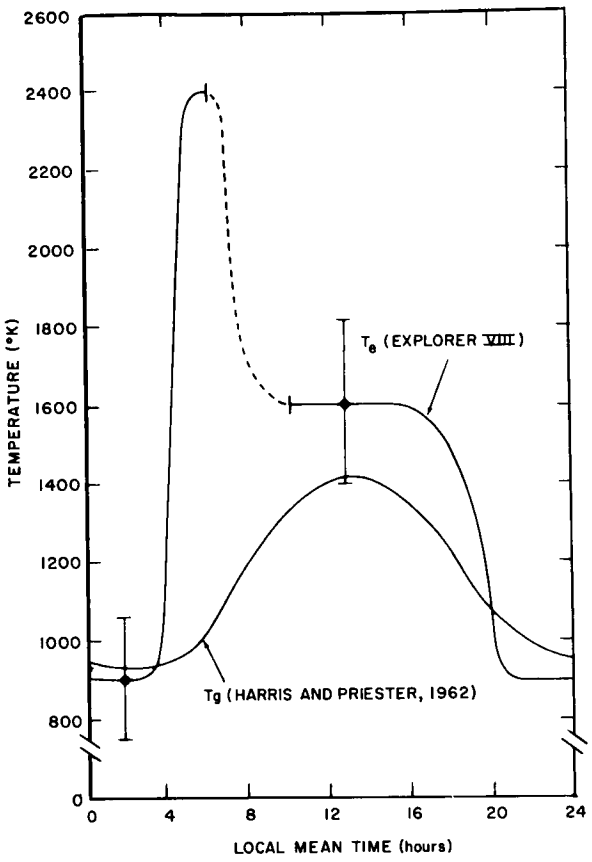


FIGURE 14.—Measured diurnal variation of electron temperature at mid-latitudes for magnetically quiet days ($A_p<15$).

figure 7 except that an electron retarding potential was applied to the collector. This particular experiment had a poorer telemetry resolution than the two-electrode experiment hence these data were not used extensively. For the times it was used the electron temperature values agreed with those from the two-electrode experiment.

The short active life of the Explorer VIII satellite together with the nature of real time telemetry transmissions severely limits our ability to separate latitude, altitude, and diurnal variations of T_e. This problem is made even more difficult because the months of November and December, 1960 were characterized by a number of rather severe solar flares, truly an unfavorable time interval for the proponents of temperature equilibrium between electrons and heavy particles. Accordingly, the presentation of these results has been limited to "quiet" days, specifically when the worldwide index (A_p) of magnetic activity was less than 15. The nature of the coverage available from the telemetry receiving network was such that the data apply principally to magnetic dips between 50° and 70°.

The average diurnal variation of T_e taken from approximately 100 passages over a telemetry station is presented in figure 14. Also shown is the estimated diurnal variation of the neutral gas temperature in the isothermal altitude region for this level of solar activity, taken from Harris and Priester (1962). There is the need for an immediate note of caution, *we have made the questionable assumption that the electron temperature is independent of altitude*, an assumption which can be examined in some detail as the discussion is developed. It is seen, firstly, that the average electron temperature taken during the 6-hour period centered around midnight is 900°K. The individual data points showed a standard deviation perhaps reflects expected daily variations in the neutral gas temperature. The altitude interval over which these nighttime data apply is 425–600 km. The rocket results of Brace and Spencer (1963) justify the assumption that T_e should be constant with altitude at the diurnal minimum. Our average nighttime value of 900°K is in excellent agreement with estimates of the neutral gas temperature for this epoch of the solar cycle, indicating no nighttime source of ionization for the stated conditions. Although our nighttime values were obtained principally for magnetic dips 50–70° N, the little data we have at other latitudes show values that are within the standard deviation at magnetic dips between 0 and 75°N. We did not observe a single value outside this deviation for $A_p<15$. Our observation that the nighttime electron temperature is consonant with the neutral gas temperature contrasts with the midnight rocket result obtained at a magnetic dip of 70°N, for low geomagnetic activity by Brace and Spencer (1963). Their value of about 1200°K was obtained in December 1961 when the estimated T_g was about 800°, thus suggesting a ratio of T_e/T_g of 1.5. We tend not to ascribe the different spaceflight results to the different experimental approaches used since, as discussed later, the Explorer VIII midday results are in

harmony with those of Brace and Spencer. They ascribe their high nighttime value to the possible existence of a local ionization source. Alouette satellite results obtained in late 1962) (Knecht and Van Zandt, 1963) show that spread-F conditions are a characteristic feature of the upper ionosphere about 70° magnetic dip. This feature also probably is related to the existence of a secondary ionization source. Willmore et al. (1962) in their preliminary analysis of Ariel satellite results (which also supposes that T_e is independent of altitude) report a latitude dependence of nighttime electron temperature with a spread that lies just outside our standard devitation. Their high latitude values are significantly above the neutral gas temperature inferred from current reference atmospheres. The most probable means of accommodating the Explorer VIII nighttime results with these other spaceflight observations is to suggest a nighttime ionization source at high magnetic dips which grows in relative importance as one approaches the year of minimum solar activity.

Turning now to the midday 6-hour period centered at 1300 LMT, the average observed electron temperature for this time interval was 1600°K with a standard deviation of 200°K. This portion of the diurnal variation was obtained at altitudes between 1000–2000 km and mostly over Australia (magnetic dip 50–70°S). We have not included in this average a few percent of the total number of observations where anomalously high values were observed. These could represent residual effects of solar flares which characterized the active life of Explorer VIII. The average value of 1600°K is within 200°K of current estimates of the neutral gas temperature at diurnal maximum for the period November–December 1960. Our average value agrees well with other spaceflight results obtained for the same latitude, similarily "quiet" ionospheric conditions and during the same epoch of the solar cycle. Hanson (1962) infers from an ion density profile taken in October 1960 that $(T_e+T_i)/2$ was 1600°K, data which apply to our altitude interval. When Hanson's result is compared with the Explorer VIII T_e result, we conclude that $T_e \approx T_i$ for this general time interval and in this altitude region. We also infer close equivalency of the charged particle and neutral gas temperatures. Assuming altitude-independent temperatures, the Explorer VIII result also is in good agreement with the value for $(T_e+T_i)/2$ of 1600°K obtained in the altitude region 350–600 km by Jackson and Bauer (1961) from an electron density profile measured on a rocket launched in April 1961. Again assuming altitude-independent temperatures, we also are in good agreement with the March 1961 rocket measurement of T_e obtained at about 350 km by Brace and Spencer (1963) who conclude: "...thermal equilibrium is normal in the quiet, daytime ionosphere at midlatitude, except in the lower F-region (approximately 150–300 km)."

We emphasize that all of the above described rocket results and the conclusion of temperature equilibrium derived therefrom were obtained at midday, at midlatitudes under quiet ionospheric conditions and for a specific epoch of the solar cycle. *Evidence is growing that to generalize the conclusion of temperature equilibrium for all latitudes and for other portions of the solar cycles would be premature.* Brace and Spencer (1963) report high electron temperatures in the auroral region during 1960. Evans (1962) reports a value for T_e/T_i of 1.6 above 200 km obtained in 1962 at magnetic dip 70°N from ground-based incoherent backsetter observations. The reported high-latitude midday T_e results from the Ariel satellite (Willmore et al., 1962) have values significantly higher than the estimate neutral gas temperature for the pertinent time interval (May, 1962). Since some of these data were obtained at the quoted Explorer VIII latitudes, there is an inference that the ratio T_e/T_i tends to increase considerably both as we approach solar minimum and/or go to high latitudes.

We consider last the most pronounced feature of the Explorer VIII diurnal T_e variation—the high electron temperatures observed in the sunrise period. These values of up to 2.5 times the estimated neutral gas temperature were obtained in the altitude interval 600–900 km at magnetic dips 50°–70°. Bowles et al. (1962) report a maximum in T_e/T_i during the sunrise period from ground-based incoherent backsetter results obtained at the geomagnetic equator. Evans (1962), on the other hand, reports that T_e/T_i reaches a maximum at noon. Hanson (1962b) and Dalgarno

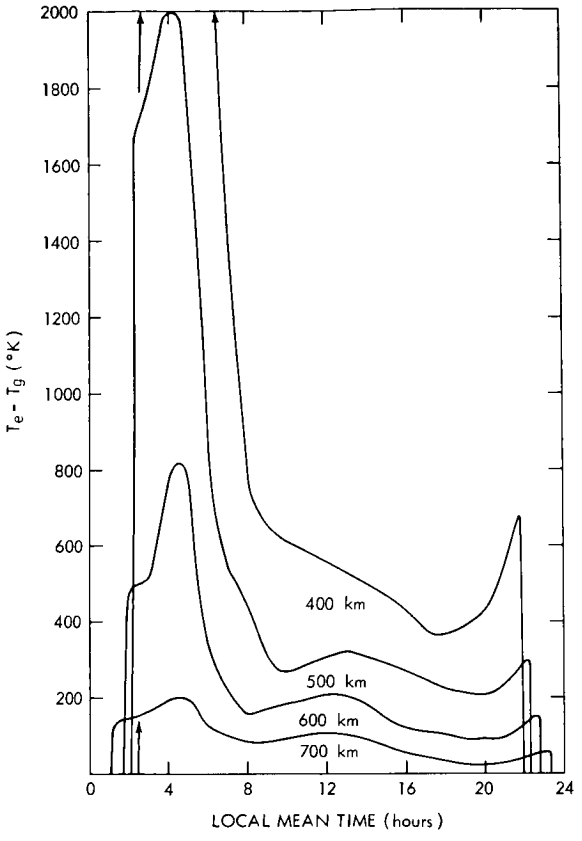


FIGURE 15.—Qualitative diurnal variation of difference between electron and ion temperatures, predicted from diurnal variation of $n(O)$ and n_e.

et al. (1962) on the assumptions that solar radiation is the only ionizing source and that the excess photoelectron energy is deposited below the F2 peak have estimated theoretically small differences in T_e and T_i in the upper ionosphere but only for a midday ionosphere. Both works relate the temperature difference $(T_e - T_i)$ for altitudes above the F2 peak to the ratio $n(O)/n_e^2$ where $n(O)$ is the number density of oxygen atoms. Dalgarno (1963) using hypothetical values for n_e has extended the theoretical estimates to suggest high electron temperatures in the sunrise period. We have applied this reasoning to a more practical case by dividing the diurnal variation of $n(O)$ given by Harris and Priester (1962) by the diurnal n_e^2 variation reported by Blumle et al. (1963) from the Alouette satellite, both variations applying to the period of late 1962. The result for different altitude increments is presented in figure 15. It shows that high electron temperatures should be expected in the sunrise period but that the effect becomes rather diffuse at altitudes above 700 km. It additionally shows that the temperature difference can increase somewhat in the sunset period. We emphasize the qualitative nature of figure 15 principally because (1) the solar ultraviolet intensity was not measured at the same time and (2) $n(O)$ is an inferred rather than a measured parameter. However if the shape of the $n(O)$ variation as a function of diurnal time is correct, we would expect T_e/T_i to be largest at sunrise and increase somewhat at sunset for the altitude interval of the Explorer VIII measurements. We do observe that T_e is a maximum at sunrise and that T_e/T_i is somewhat larger at sunset than at midday. The sunset data shown in figure 13 were obtained between 600–900 km. We did not include in this averaging several higher altitude observations where T_e was higher than the average curve. Hanson (1962b) has suggested the possibility that some photoelectrons can diffuse along magnetic field lines and deposit their energy at higher altitudes. This would be critically dependent on the relationship between magnetic dip and the solar zenith angle. It is possible that the higher T_e values observed above 900 km at sunset are related to this hypothesis.

REFERENCES

BAUER, S. J., "Helium Belt in the Upper Atmosphere," *Nature*, **197**, 36–37, Jan. 5, 1963.

BLUMLE, L. J., FITZENREITER, R. J., BAUER, S. J., and JACKSON, J. E., "Diurnal Variations of the Topside Ionosphere at Mid-latitudes," presentation at 44th annual meeting of the American Geophysical Union, Washington, D.C., Apr. 18, 1963.

BOURDEAU, R. E., "Ionospheric Results With Sounding Rockets and the Explorer VIII Satellite," in: *Space Research II; Proc. 2d Internat. Space Sci. Symp., Florence, Apr.* 10–14, 1961, Amsterdam: North-Holland Publishing Co., 1961, pp. 554–573.

———, DONLEY, J. L., SERBU, G. P., and WHIPPLE, E. C., Jr., "Measurements of Sheath Currents and Equilibrium Potential on the Explorer VIII Satellite," *J. Astronaut. Sci.*, **8**(3): 65–73, 1961.

———, WHIPPLE, E. C., Jr., DONLEY, J. L., and BAUER, S. J., "Experimental Evidence for the Presence of Helium Ions Based on Explorer VIII Satellite Data," *J. Geophys. Res.*, **67**(2): 467–476, February 1962.

BOWLES, K. L., OCHS, G. R., and GREEN, J. L., "On the Absolute Intensity of Incoherent Scatter Echoes From the Ionosphere," *J. Res. Nat. Bur. Stand. (D. Rad. Prop.)* **66D**(4): 395–407, July–August 1962.

BRACE, L. H., SPENCER, N. W., and CORRIGNAN, G. R., "Ionosphere Electron Temperature Measurements and Their Implications," *J. Geophys. Res.*, **68**(19): 5397–5412, Oct. 1, 1963.

DONLEY, J. L., "Experimental Evidence for a Low Ion-Transition Altitude in the Upper Nighttime Ionosphere," *J. Geophys. Res.*, **68**(7): 2058–2060, Apr. 1, 1963.

EVANS, J. V., "Diurnal Variation of the Temperature of the F Region," *J. Geophys. Res.*, **67**(12): 4914–4920, November 1962.

GRINGAUZ, K. I., and ZELIKMAN, M. KH., "Measurement of the Positive Ion Density Along the Orbit of an Artificial Earth Satellite," *Uspekhi Fiz. Nauk*, **63**(1b): 239–252, 1957 (in Russian).

HANSON, W. B., "Upper Atmosphere Helium Ions," *J. Geophys. Res.*, **67**(1): 183–188, January 1962.

HARRIS, I., and PRIESTER, W., "Theoretical Models for the Solar-Cycle Variation of the Upper Atmosphere," *J. Geophys. Res.*, **67**(12): 4585–4591, November 1962.

HINTEREGGER, H. E., DAMON, K. R., and HALL, L. H., "Analysis of Photoelectrons From Solar Extreme Ultraviolet," *J. Geophys. Res.*, **64**(8): 961–969, August 1959.

JACKSON, J. E., and BAUER, S. J. "Rocket Measurement of a Day-time Electron-Density Profile Up to 620 Kilometers," *J. Geophys. Res.*, **66**(9): 3055–3057, September 1961.

KNECHT, R. W., and VAN ZANDT, T. E., "Some Early Results From the Ionospheric Topside Sounder Satellite," *Nature*, **197**(4868): 641–644, Feb. 16, 1963.

KRASSOVSKY, V. I., "Exploration of the Upper Atmosphere With the Help of the Third Soviet Sputnik," *Proc. I.R.E.*, **47**(2): 289–296, February 1959.

SCHULZ, G. J., and BROWN, S. C., "Microwave Study of Positive Ion Collection by Probes," *Phys. Res.*, **98**(6): 1642–1649, June 15, 1955.

SERBU, G. P., BOURDEAU, R. E., and DONLEY, J. L., "Electron Temperature Measurements on the Explorer VIII Satellite," *J. Geophys. Res.*, **66**(12): 4313–4315, 1961.

WHALE, H. A., "The Excitation of Electroacoustic Waves by Antennas in the Ionosphere," *J. Geophys. Res.*, **68**(2): 415–422, Jan. 15, 1963.

WHIPPLE, E. C., Jr., "The Ion-Trap Results in 'Exploration of the Upper Atmosphere With the Help of the Third Soviet Sputnik'," *Proc. I.R.E.*, **47**(11): 2023–2024, November 1959.

IONOSPHERE ELECTRON TEMPERATURE MEASUREMENTS AND THEIR IMPLICATIONS

L. H. BRACE AND N. W. SPENCER

Goddard Space Flight Center

G. R. CARIGNAN

Space Physics Research Laboratory, Department of Electrical Engineering, University of Michigan

Electron temperature and ion density data, reported earlier in preliminary form, are reviewed in terms of their consistency with recent theories regarding thermal equilibrium in the daytime ionosphere. On the basis of these daytime data and the ionosphere theory, we conclude that solar extreme ultraviolet radiation (EUV) is the dominant heat source in the quiet midlatitude ionosphere. This conclusion is suggested by the agreement between the electron temperature profiles measured under these conditions and theoretically predicted profiles based on EUV radiation as the only heat source. Measurements in the disturbed midlatitude and auroral zone ionospheres suggest that, although EUV remains the dominant heat source for electrons in the F_1 region, an additional heat source is effective in the E and F_2 regions. Its effect is to extend the region of thermal nonequilibrium both upward into the upper F region and downward into the E region. The results of a more recent flight in the nighttime ionosphere are presented and discussed. A horizontal gradient in electron temperature found in the nighttime ionosphere is believed too abrupt to permit a similar gradient in the gas temperature and thus indicates that small but variable degrees of thermal nonequilibrium can exist in the nighttime ionosphere. This is believed to be because of particle fluxes whose heating effects are enhanced by the low values of electron density existing in the nighttime F region.

INTRODUCTION

The solar energy incident upon the earth's atmosphere in various forms (extreme ultraviolet, X ray, corpuscular) ionizes neutral particles in the thermosphere and thereby generates electrons having energies of tens of electron volts. Most of this energy is lost by inelastic collisions with the neutral particles, but the significant amount remaining after the last inelastic collision (approximately 2 ev) is rapidly shared by elastic collisions with the ambient electrons. Since the ambient electrons represent only a minority constituent of the ionosphere, this tends to raise the electron temperature above the gas temperature and produces a degree of thermal nonequilibrium in the altitude region where these processes are important. On the basis of these considerations, several investigators (Hanson and Johnson, 1961; Hanson, 1962; Dalgarno et al., 1962) have calculated the degree of nonequilibrium to be expected because of solar extreme ultraviolet (EUV) alone. Though their models of electron temperature differ in detail, all show that the absorption of EUV should cause the electron temperature to exceed the gas temperature between approximately 150 and 350 km in the quiet midday ionosphere. Their results, to be shown later, indicate that the T_e/T_g ratio reaches values of 1.5 to 2 (between 150 and 350 km), depending on the actual energy transfer mechanisms occurring there, and that thermal equilibrium is probably reestablished above 350 km. It is important to note that these authors have not considered the

effects of other heat sources such as solar X radiation and corpuscular radiation which may be of great importance in the disturbed or higher-latitude ionosphere and are probably significant even under quiet conditions at midlatitudes.

Experimental data concerning thermal equilibrium are conflicting. Evans (1962) has recently reported radar backscatter measurements indicating that the electron temperature above 300 km at midday normally exceeds the ion (and presumably neutral) temperature by a factor of 1.6. Hanson and McKibbin (1961), Jackson and Bauer (1961), Brace et al. (1962), and others have, by assuming thermal equilibrium, used the scale-height concept to interpret rocket measurements of electron or ion concentration in the upper F region in terms of atmospheric temperature. The consistency of these data with temperatures derived indirectly from satellite drag studies has been cited as partial justification for the assumption of thermal equilibrium [Bauer and Bourdeau, 1962].

In a recent paper Spencer et al. (1962) reported the data from a series of four daytime direct measurements of electron temperature in the E and F regions. These data confirmed that the electron temperature exceeds the generally accepted values of neutral particle temperature in the F_1 region, as had been predicted earlier by Hanson and Johnson (1961) and later by Hanson (1962) and Dalgarno et al. (1962). However, only the single flight of this series that was made under quiet ionospheric conditions at Wallops Island showed the negative temperature gradient in the F_2 region that would be required to permit a return to thermal equilibrium there. The other three measurements of the series, including one in a disturbed ionosphere above Wallops Island and two in the auroral zone ionosphere (Fort Churchill), indicated that the region of nonequilibrium extended to more than 400 km with generally increasing temperature above the F_2 maximum. These data tend to support Evans' results.

Unfortunately there exist too few measurements of electron temperature in the 150- to 450-km altitude range to permit us to decide which electron temperature profile is normal: the "quiet" ionosphere with its apparently well-behaved region of thermal nonequilibrium between 150 and 350 km, or the "disturbed" ionosphere in which nonequilibrium extends down into the E region and well into the upper F region.

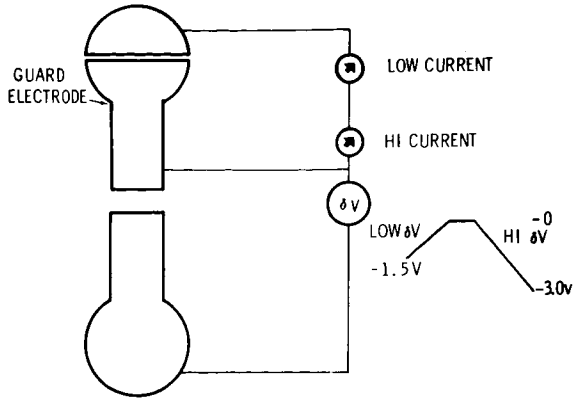


FIGURE 1.—The measurement system employed in the Dumbbell experiment. A modified sawtooth voltage is applied between opposite ends of the symmetrical system so as to drive the insulated hemisphere and its associated guard negative with respect to the reference end of the instrument.

Since our earlier paper, in which preliminary data from NASA 6.03 and NASA 6.04 were presented, the analysis of data from these Wallops Island flights has been completed, enabling us to confirm our earlier conclusions about thermal equilibrium in the daytime ionosphere. In addition, a subsequent nighttime measurement at Wallops Island has been carried out and the data analyzed. It is the purpose of this paper to present final data from the earlier flights, as well as data from the more recent nighttime flight, and to interpret the results of the entire flight series in terms of their consistency with certain aspects of ionosphere theory that have recently been reported. Emphasis will be on the nighttime data that has not been published previously.

THE MEASUREMENT TECHNIQUE

The direct measurement technique employs a dumbbell-shaped symmetrical bipolar probe that is ejected from the rocket vehicle at the base of the ionosphere, rises to several hundred kilometers, and returns to earth. While it is in the ionosphere, a sawtooth voltage is applied to a hemispherical collector, and the resulting

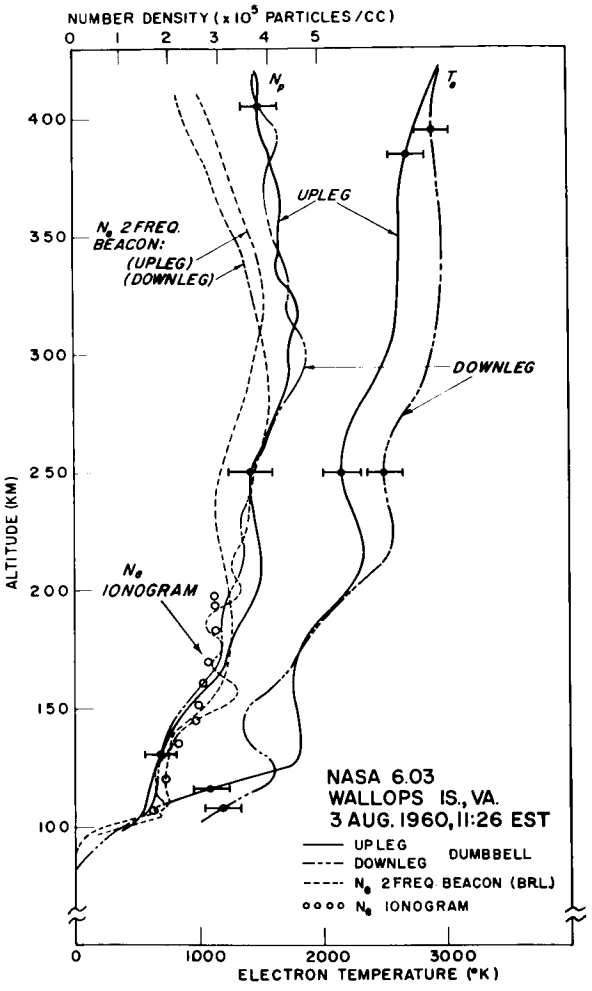


FIGURE 2.—Final ionospheric data from NASA 6.03 (disturbed day). The electron temperature T_e and the positive ion density N_p derived from the Dumbbell volt-ampere characteristics are shown. The error flags represent both spread in the data and possible errors in the theory. Also shown are the electron densities N_e measured by the two-frequency beacon carried with the rocket (courtesy W. Berning) and those derived from ground-based ionosonde records.

currents are measured and telemetered to receiving stations where the data are recorded for later analysis in terms of the electron temperature and positive ion density, point by point along the flight path. Figure 1 shows the measurement circuit. The magnitude of the recorded current is proportional to the local ion density N_p, and the shape or curvature of each volt-ampere characteristic is related to temperature of the ambient electrons T_e. Details of the measurement theory and the methods of data analysis are documented elsewhere (Spencer et al., 1962; Hoegy and Brace, 1961; Brace, 1962).

FINAL DATA FROM DAYTIME WALLOPS ISLAND FLIGHTS

The complete data from NASA flights 6.03 and 6.04 (previously reported in preliminary form) are shown in figures 2 and 3, respectively. Also shown are the electron density profiles derived from a two-frequency beacon experiment (courtesy of Warren Berning, BRL), which remained with the vehicle after Dumbbell ejection, and the electron densities derived from the ground-based ionosonde records (courtesy of J. W. Wright,

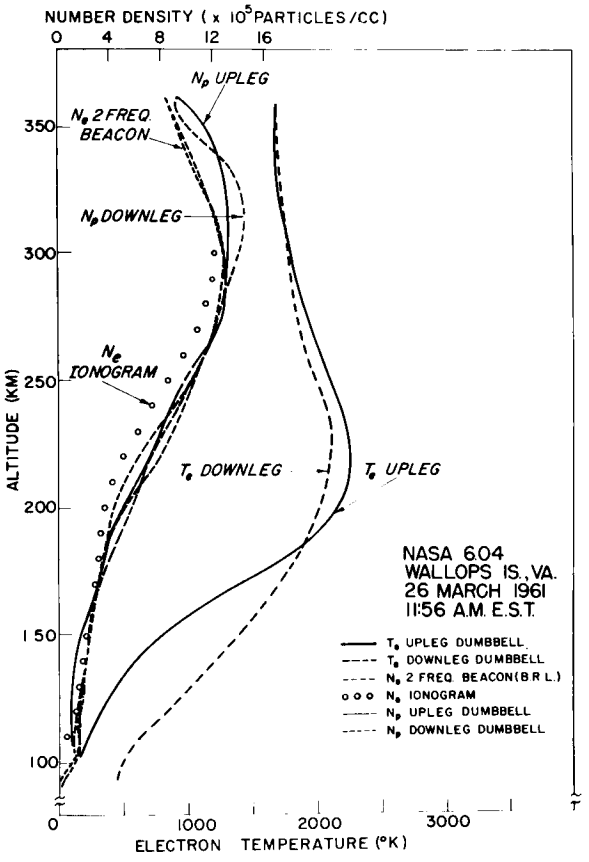


FIGURE 3.—Final ionospheric data from NASA 6.04 (quiet day). The electron temperature T_e and the positive ion density N_p derived from the Dumbbell data are shown with the N_e values from the two-frequency beacon and ionosonde.

NBS). The combined effects of spread in the individual data points and our estimate of the uncertainty in the theory are indicated by the error flags. The uncertainty in the theory tends to be systematic and does not change the shape of the profiles. These midday measurements were carried out about 8 months apart at the same midlatitude launch site (38°N); however, the ionospheric conditions at the times of launch were quite different. Magnetic storms were reported by several observers (Lincoln, 1961) during the days preceding the launch of NASA 6.03, July 29 to August 1, 1960. Although the launch date was not listed as disturbed, the severe depression of the F_2 maximum, readily apparent in both the electron and ion density profiles, undoubtedly was part of the so-called "after effect" of an ionospheric storm (Ratcliffe and Weekes, 1960). The period immediately preceding the launching of NASA 6.04, however, was unusually quiet magnetically. In the discussion, the data from these flights will be considered in terms of the degree of disturbance of the ionosphere.

NIGHTTIME MIDLATITUDE IONOSPHERIC DATA

The most recent launching of the Dumbbell series was carried out at Wallops Island shortly before midnight on December 21, 1961. At an altitude of 86 km the clamshell nose cone opened and ejected the instrument into a region having

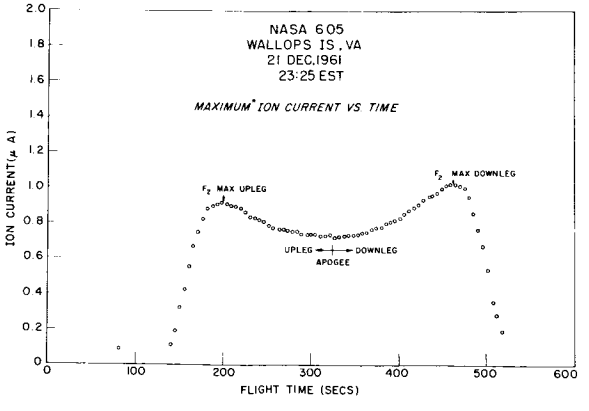


FIGURE 4.—Ion current data from NASA 6.05. Maximum ion current refers to the hemisphere current measured at the maximum applied voltage (−3 v).

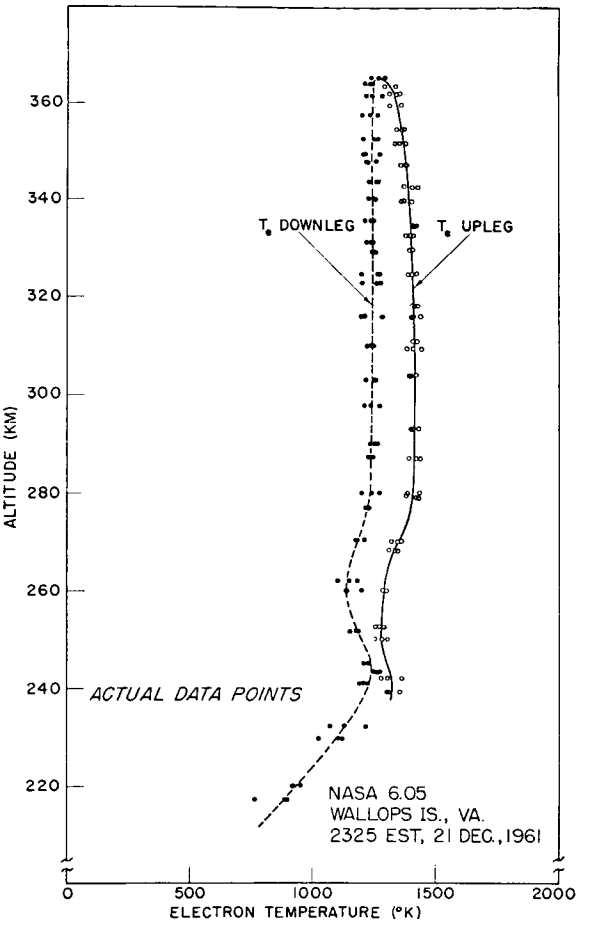


FIGURE 5.—Actual electron temperature data points from NASA 6.05 flight. At selected altitude intervals, three consecutive current characteristics were analyzed for electron temperature. The open and closed circles represent the upleg and downleg values, respectively.

an ion density N_p below its limit of resolution ($N_p < 1 \times 10^3$ ions/cc). A small nighttime E region was detected between 95 and 105 km above which the current again fell below the detectable limit. F-region currents were first detected above 204 km and were well resolved through apogee of the flight, as shown in figure 4, which is a plot of the ion current versus flight time. The times of F_2 passage are clearly evident at 200 sec and 460 sec. These values of ion current were used in the appropriate equations (Hoegy and Brace, 1961) to calculate the ion density values shown in figure 6. The E-region values of N_p were near the limit of detection but are estimated to be accurate within a factor of 2. The accuracy at higher

TABLE I.—*Tabulated Dumbbell Probe Data*

Altitude (km)	NASA 6.03				NASA 6.04				NASA 6.05			
	Up-Leg		Down-Leg		Up-Leg		Down-Leg		Up-Leg		Down-Leg	
	T_e (°K)	N_p (10^5/cc)	T_e (°K)	N_p (10^5/cc)	T_e (°K)	N_p (10^5/cc)	T_e (°K)	N_p (10^5/cc)	T_e (°K)	N_p (10^5/cc)	T_e (°K)	N_p (10^5/cc)
90	------	--------	------	0.35	------	0.31	------	0.55				
100	------	1.05	------	0.93	------	1.19	485	1.38				
110	730	1.45	1260	1.56	245	1.62	615	1.70				
120	1360	1.57	1590	1.70	330	0.94	840	1.55				
130	1810	1.72	1560	1.73	445	1.00	1025	1.85				
140	1825	1.95	1370	1.88	610	1.30	1210	2.10				
150	1795	2.28	1400	2.17	800	1.70	1380	2.40				
160	1770	2.75	1620	2.57	1070	2.45	1545	2.60				
170	1790	3.08	1780	2.93	1410	3.00	1695	3.25				
180	1880	3.22	1910	3.00	1730	3.60	1840	4.00				
190	2040	3.50	2060	3.02	2000	4.05	1950	4.80	------	--------	------	0.02
200	2220	3.72	2250	3.20	2160	4.95	2020	5.65	------	--------	------	0.18
210	2320	3.81	2470	3.41	2240	6.00	2075	6.40	------	0.25	------	0.47
220	2340	3.75	2580	3.45	2270	7.15	2110	7.15	------	0.98	935	0.98
230	2295	3.71	2580	3.40	2250	8.00	2115	8.30	------	1.43	1100	2.30
240	2215	3.62	2540	3.57	2190	8.80	2080	9.35	1320	2.70	1225	4.05
250	2180	3.62	2520	3.62	2120	9.80	2000	10.30	1280	4.40	1200	5.45
260	2185	3.82	2550	3.87	2045	11.00	1930	11.35	1290	5.80	1140	6.60
270	2240	4.02	2610	4.05	1970	12.25	1875	12.20	1355	6.95	1185	7.90
280	2340	4.21	2745	4.30	1900	12.90	1840	12.70	1410	7.75	1237	8.75
290	2460	4.38	2840	4.61	1845	13.00	1810	13.30	1416	8.20	1240	9.20
300	2545	4.40	2890	4.74	1810	13.10	1785	14.20	1419	8.55	1243	9.44
310	2600	4.48	2940	4.60	1780	13.25	1765	14.80	1420	8.75	1245	9.49
320	2625	4.58	2965	4.36	1750	13.20	1725	14.95	1412	8.70	1249	9.40
330	2640	4.33	2985	4.40	1725	12.85	1700	14.20	1409	8.50	1249	9.20
340	2640	4.16	2995	4.28	1710	12.30	1690	12.10	1400	8.49	1248	8.97
350	2640	4.20	2995	4.08	1700	11.50	1690	10.20	1377	8.63	1244	8.82
360	2650	4.24	2980	3.98	1685	9.80	1685	9.25	1345	8.85	1249	8.85
370	2655	4.22	2965	3.92								
380	2685	4.07	2950	4.00								
390	2740	3.94	2925	4.18								
400	2810	3.80	2915	4.02								
410	2880	3.82	2940	3.74								
420	2965	3.77	2990	3.73								

altitudes is indicated by the error flags shown. Also shown are the electron density values N_e derived from an ionogram recorded at Wallops Island during the flight. The low electron density made the ionogram analysis difficult, particularly in converting from virtual height to actual height. This is evident from the difference in the altitude of the F_2 maximum indicated by the ionosonde data (350 km) and by the probe data. The actual altitude of the maximum (310 km) is best derived from the ion current data shown in figure 4. Note that the maximums in the N_p profiles are less pronounced than those in the current data owing to the correction for probe velocity in the density calculation.

Above 240 km on ascent and 220 km on descent

the charge densities, and therefore the currents, were large enough to allow precise extraction of the electron current from the volt-ampere characteristics and thus permit analysis of the data in terms of the electron temperature of the surrounding plasma. Figure 5 shows the actual temperature data points, each representing a single probe characteristic, used to arrive at the T_e profile shown in figure 6.

The data from NASA 6.03, 6.04, and 6.05 are tabulated at 10-km intervals in table 1.

DISCUSSION

The Quiet Ionosphere (NASA 6.04)

It is interesting to compare the electron temperatures measured in the quiet ionosphere with those calculated by Hanson (1962) and Dalgarno

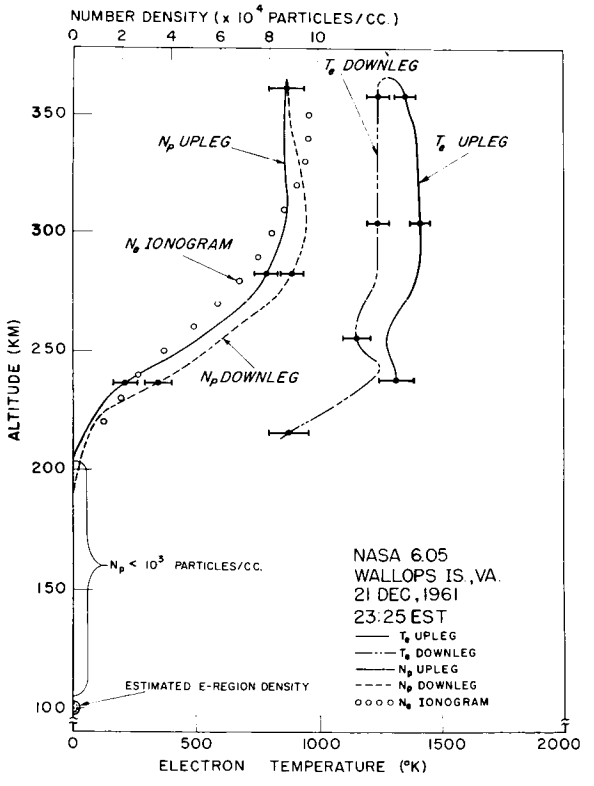


FIGURE 6.—Ionosphere data from NASA 6.05 (nighttime). T_e and N_p values from the Dumbbell are plotted with the ionosonde values of N_e. Between 105 and about 200 km the ion currents were too small to resolve. The E-region values of N_p were 4×10^3/cc on the upleg and 2×10^3/cc on the downleg.

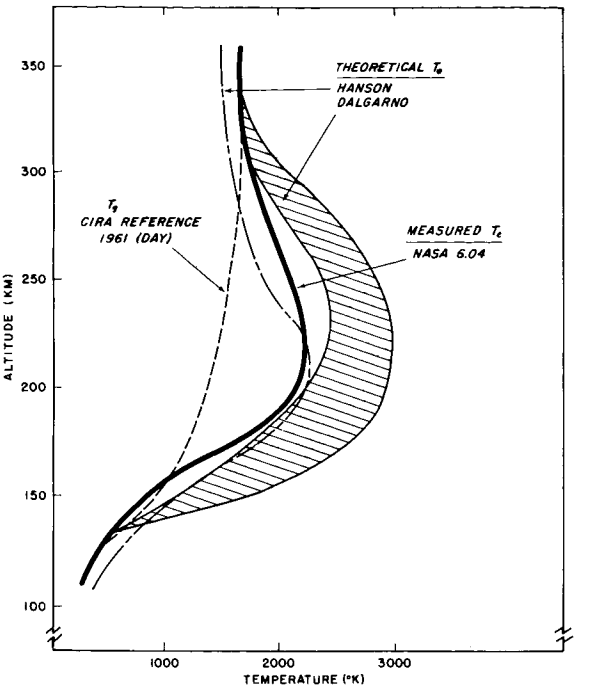


FIGURE 7.—Comparison of measured and theoretical values of electron temperature in the quiet daytime ionosphere.

et al. (1962) (figure 7). The CIRA 1961 average maximum gas temperature profile is shown for reference. As was mentioned earlier, both authors based their computations on rocket measurements of solar EUV flux and the important energy transfer mechanisms, and both predicted a region of thermal nonequilibrium in the F region similar to that evident in the electron temperature data from this flight in the quiet ionosphere. Dalgarno's temperatures are shown as a range of values (hatched area) corresponding to his selection of combinations of the two most probable energy storage mechanisms, namely, vibrational excitation of molecular nitrogen and excitation to metastable states of atomic oxygen. This stored energy is later returned in part to the electrons, thereby further increasing their temperature above that of the neutral particles.

The agreement between these theoretical and experimental temperatures demonstrates that solar EUV alone is an adequate energy source to produce the observed electron heating in the quiet daytime ionosphere at midlatitudes. If other heat sources are present under these conditions, their

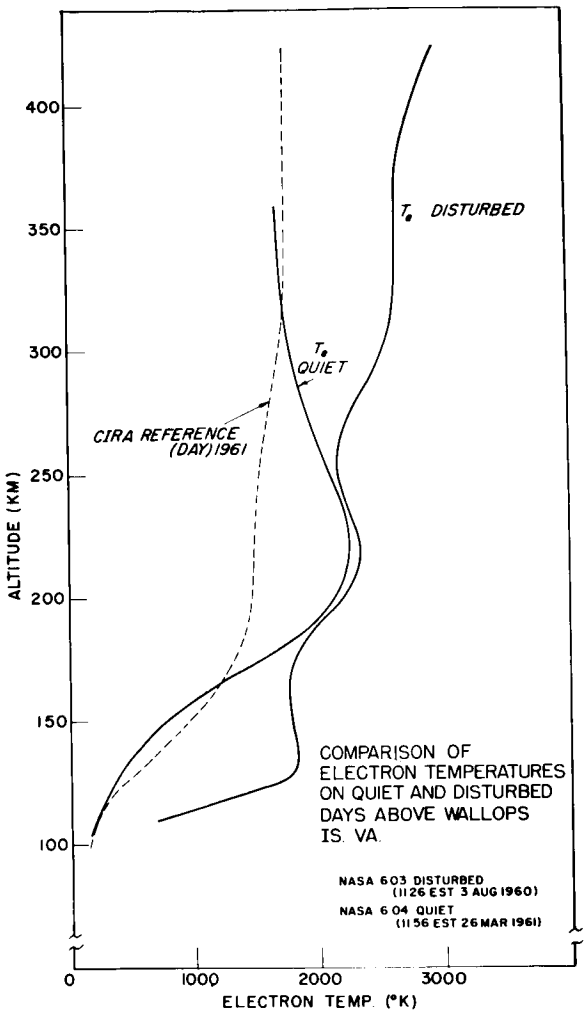


FIGURE 8.—Comparison of electron temperatures in the quiet and disturbed midlatitude ionosphere with the reference gas temperature.

effects must be masked by the uncertainty of the neutral temperature existing at the time of the electron temperature measurement.

The Disturbed Ionosphere (NASA 6.03)

The clearly different character of the disturbed ionosphere temperature profile becomes evident when it is compared with the quiet ionosphere data (fig. 8). The gross effect is an increased electron temperature in both the E and F_2 regions with apparently little effect in the F_1 region. The steep temperature gradient between 105 and 125 km, accompanied by a higher than normal E-region density (fig. 2), is evidence of an additional energy input at these altitudes. It is not yet clear, however, whether the higher temperatures in the F_2 region were caused by an increased heat input at these altitudes or by sharing the existing EUV energy among the reduced number of ambient electrons there under these conditions.

The Nighttime Ionosphere (NASA 6.05)

Possibly the most unexpected feature of the nighttime temperature profiles (fig. 6) is the temperature difference of the regions traversed on ascent and descent. Although the downleg values are in agreement with the 1961 CIRA reference gas temperatures, they are some 5 to 15 per cent lower than the upleg values. Both profiles are essentially isothermal with altitude in the F_2 region, in agreement with heat conduction theory (Nicolet, 1960). Thus we interpret the change in temperature occurring near apogee as evidence of a horizontal gradient traversed by the instrument in its largely horizontal path through the F_2 region.

(It is important to note that the nature of the Dumbbell experiment permits us to make detailed comparison between upleg and downleg data with some assurance. This is true for two reasons; first, the ejection of the device carries it sufficiently far ahead of the rocket, its exhaust

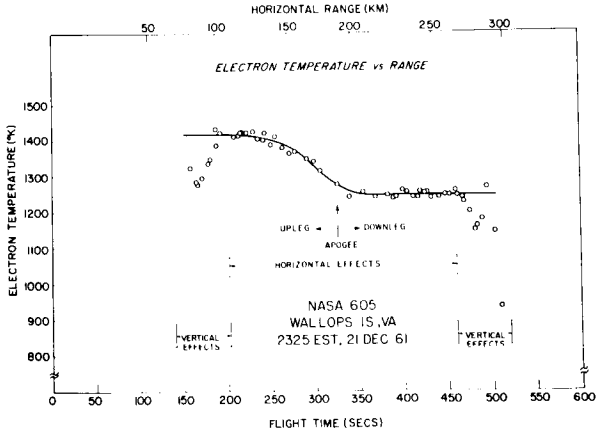


FIGURE 9.—Plot of electron temperature versus horizontal range. This reveals the temperature gradient, encountered near apogee of NASA 6.05, as a horizontal gradient.

gases, and its charged particle wake to eliminate possible perturbing effects on the collected currents; second, the tumbling motion of the device permits the collectors to experience identical orientation and velocity effects at equivalent altitudes on ascent and descent. Thus the ejection concept and tumbling motion permit greater confidence that observed differences in upleg and downleg data are due to natural atmospheric variations rather than experiment contamination or perturbations in the mode of current collection. Periodic inflight calibration of the current channels, using the sawtooth voltage and a known resistance, provide added assurance that any changes in measured parameters are not caused by changes in the electrical characteristics of the measurement system.)

To investigate further the nature of the gradient, the temperature data points were replotted versus flight time or horizontal range (figure 9). Each point represents an average of the groups of three points of figure 5. Here it is clear that the entire change in temperature occurred over a horizontal range of 75 km. Since this corresponded to only a 25-km range in altitude, the gradient must have been primarily horizontal in nature.

A certain amount of horizontal gradient will normally result from the diurnal temperature variation experienced by the instrument as its eastward horizontal motion and the elapsed time of flight combine to change its local time by approximately 30 minutes. However, currently held concepts about the nature of the normal diurnal variation of neutral particle temperature are not compatible with the degree of temperature decrease observed. For example, the Harris and Priester (1962) model (fig. 10) suggests that a decrease in gas temperature of only a few percent should be expected in the half-hour period shortly before midnight. The upleg and downleg values of NASA 6.05 and the apogee value of NASA 6.04 are plotted for comparison with the theoretical model for a 10.7-cm flux of 150×10^{-22} W/(m^2 c/s), which corresponds generally to the levels of solar activity in 1960 and 1961.

If we assume that thermal equilibrium existed at apogee of the NASA 6.04 daytime flight, as the approach of the temperature profile to an

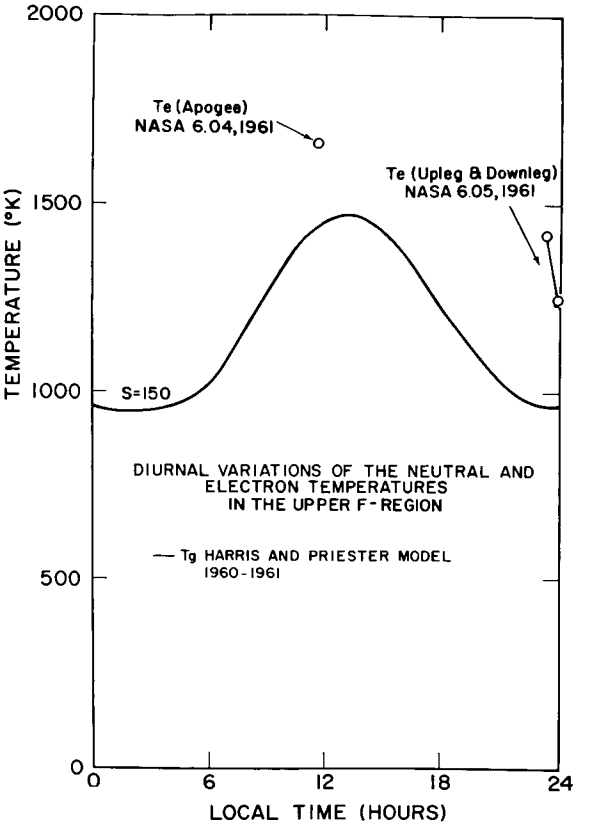


FIGURE 10.—Comparison of neutral temperatures [Harris and Priester, 1962] and electron temperature diurnal variation.

isothermal suggests, and accept the current belief that thermal equilibrium exists at night (NASA 6.05), we are tempted to infer from figure 10 that the actual gas temperatures are normally somewhat higher than the Harris and Priester models suggest, particularly at night.

An alternate conclusion is that thermal equilibrium may not invariably exist in the nighttime ionosphere. It is conceivable, for example, that particle fluxes traveling along geomagnetic field lines could selectively heat the electrons and raise their characteristic temperature somewhat above that of the neutral particles. A similar degree of particulate heating would not be as noticeable in the quiet daytime ionosphere (figure 3), since it would be masked by solar EUV heating. If the particle flux were inhomogeneous and time-varying, different degrees of nonequilibrium would exist over different parts of the earth, and hori-horizontal gradients in the electron temperature

would result. Furthermore, the geomagnetic field at middle and high latitudes inhibits the horizontal conduction of energy between electrons in neighboring regions; thus horizontal temperature gradients, once established, would be maintained by a relatively small heat input.

Hanson and Johnson (1961) have shown that, in the region above 300 km where the energy of the thermal electrons is coupled primarily to the ions, the difference between the electron and gas temperature is given approximately by the expression

$$T_e - T_g = 2.1 \times 10^6 \, Q T_e^{3/2} / n_e^2$$

where Q is the input energy in the units of ev/cm^3/sec. Clearly the horizontal variation of T_e depends not only on the variations in Q but on the variations in T_g and n_e as well. Note that the degree of nonequilibrium is particularly sensitive to n_e, since it enters as the square. The 10 percent difference in the ascent and descent values of charge density, evident in figure 6, would alone induce a 20 percent difference in electron temperature, somewhat greater than was actually observed. It is significant that the magnitude of n_e observed on both ascent and descent of NASA 6.05 were well below the normal midnight values. This would contribute to a larger value of $(T_e - T_g)$ if an energy flux were present.

A second unexpected feature of the nighttime temperature profiles is the "S" shaped variation between 280 and 220 km for which no geophysical explanation is offered. The currents were well resolved through most of this altitude region, and we are not aware of any instrumental errors that could produce this effect.

Comparison of Midlatitude and Auroral Zone Electron Temperatures

Other significant aspects of ionospheric behavior become apparent when the Wallops Island data are plotted with the temperature data from two flights at Fort Churchill reported earlier (figure 11). Latitude effects are apparent in the comparison of the Churchill (59°) data with the daytime Wallops Island (38°) data, and the solar-cycle variation is evident in the generally decreasing temperatures found on successive daytime flights.

Latitude effects

The consistently higher electron temperature observed in the auroral zone, particularly in the E and F_2 regions, and the gross dissimilarity of the auroral zone and the quiet midlatitude profiles are interpreted as evident of a significant additional energy source that may be more important than solar EUV in these regions of the auroral zone ionosphere. The highly variable nature of the NASA 6.01 temperature data is attributed to the spread F condition that existed.

An interesting similarity in all the daytime profiles is the tendency for the temperature in the F_1 region (approximately 200 km) to be essentially identical under a variety of ionospheric conditions at both latitudes (59°N and 38°N). This is

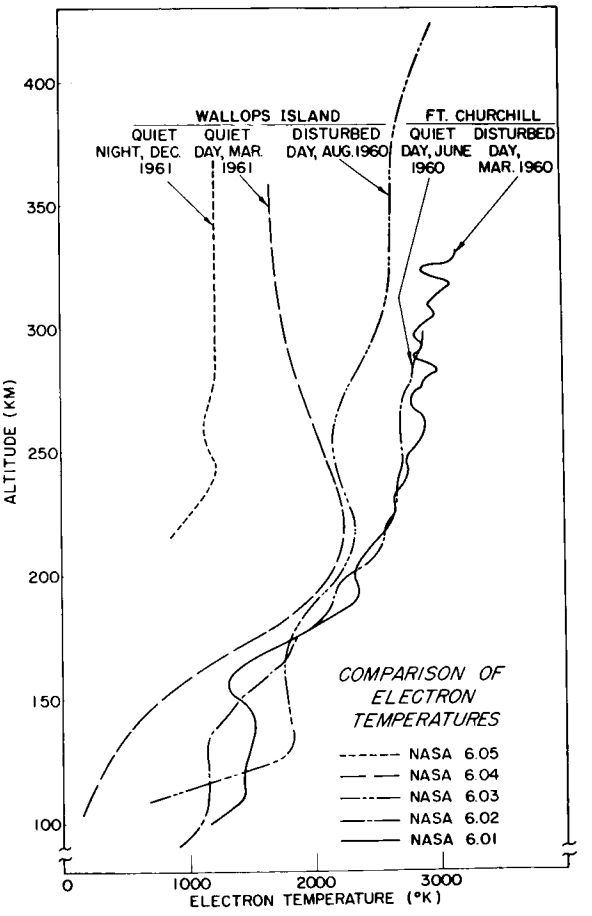


FIGURE 11.—Composite of electron temperature from the five Dumbbell flights of the NASA series carried out between March 1960 and December 1961.

in the altitude range where solar EUV absorption is greatest and is considered evidence that EUV remains the most important source of energy in the F_1 region, even under magnetically disturbed and auroral zone conditions.

Solar activity.

The effect of decreasing solar activity is apparent in the monotonically decreasing electron temperature at F_1-region altitudes on successive flights as the sunspot cycle proceeds toward its minimum. The fact that this correlation does not carry over to the E-region temperatures may be related to other heating mechanisms which become important in this region at higher latitudes and under magnetically disturbed conditions.

CONCLUSIONS

We conclude on the basis of the data presented here and current ionosphere theory that:

1. Solar extreme ultraviolet flux is the dominant heat source for the quiet daytime F region at midlatitudes (50° magnetic latitude).
2. An additional heat source, probably corpuscular, is effective in heating electrons in the E and possibly the F_2 regions of the auroral zone and disturbed midlatitude ionosphere. However, the constant electron temperature at 200 km, in both the auroral and midlatitude ionosphere, shows that EUV remains the dominant heat source in the F_1 region.
3. The "after effects" of a magnetic storm (the decreased F_2-region density and the increased F_2- and E-region electron temperature) continue for at least a day after the magnetic indices have returned to normal.
4. Particle fluxes may selectively heat the electrons in the nighttime F region and produce small but variable degrees of thermal nonequilibrium.

The evidence for conclusion 1 is the agreement between experimental electron temperatures (NASA 6.04) and theoretical models of electron temperature based exclusively on solar EUV heating. The experimental evidence for conclusion 2 is shown in figure 11, and conclusion 3 is based on figures 2 and 8. Figures 9 and 10 are considered evidence for conclusion 4.

Perhaps investigators, more familiar with particular aspects of ionospheric behavior, will uncover relationships in these data that have escaped us or will suggest alternate interpretations of the effects noted here.

It is not yet clear from these and other reported ionospheric data that the electron temperature in the ionosphere is related in any simple way to the gas temperature, except possibly under very quiet conditions. Though the electron temperature profiles from the single flight in the quiet midlatitude ionosphere appear to validate the theoretical approaches of Hanson (1962) and Dalgarno et al. (1962) for the quiet ionosphere, the data from the three other daytime flights demand other heat sources and lead us to suspect that a quiet ionosphere may have been so rare in the 1960–61 period as to be considered anomalous.

Since it is the difference between the neutral particle and electron temperatures that permits transfer of much of the solar energy into the high neutral atmosphere, we would expect the degree of thermal nonequilibrium to be strongly dependent on the solar conditions and the relative contribution of the various forms of energy input at particular latitudes and times of day. Thus the electron temperature can be expected to be a sensitive indicator of the kinds of physical processes occurring in the thermosphere. Additional measurements which are planned will be useful in evaluating new theories being evolved to describe the regions and degree of thermal nonequilibrium in the ionosphere for a wider range of conditions.

ACKNOWLEDGMENTS

We are indebted to W. W. Berning of the Ballistic Research Laboratory for the two-frequency beacon values of electron density for NASA flights 6.03 and 6.04, and to J. W. Wright of the National Bureau of Standards for the ionosonde values of electron density for all flights.

The series of measurements reported here and the development of the theory of the Dumbbell probe were carried out at the University of Michigan with the support of the National Aeronautics and Space Administration (contract

NASw-139). Early developmental work during the International Geophysical Year was supported by the Air Force Cambridge Research Laboratory and the Ballistic Research Laboratory.

REFERENCES

BAUER, S. J., and BOURDEAU, R. E., Upper atmosphere temperatures derived from charged particle observations, *J. Atmospheric Sci.*, **19,** 218, 1962.

BRACE, L. H., The dumbbell ionosphere probe; ionosphere data, *Sci. Rept. JS-3, Univ. Mich. ORA Rept.* 03599–11-*F*, September 1962.

———, SPENCER, N. W., CARIGNAN, G. R., and HOEGY, W., Paper presented at the URSI–IRE meeting, Washington, D.C., May 1962.

DALGARNO, A., MCELROY, M. F., and MOFFETT, R. J., Planetary aeronomy, 6, Electron temperatures in the ionosphere, *Geophys. Corp. Am. Tech. Rept.* 62–11–*N*, November 1962.

EVANS, J. V., Diurnal temperature variations of the *F* region, *J. Geophys. Res.*, **67,** 4914–4920, 1962.

HANSON, W. B., Electron temperatures in the upper atmosphere, *Space Research, Proc. Intern. Space Sci. Symp., 3rd, Washington,* 1962, edited by W. Priester, pp. 282–302, North-Holland Publishing Co., Amsterdam, 1962.

———, and JOHNSON, F. S., Electron temperatures in the ionosphere, paper presented at 10th International Astrophysical Colloquium, Liege, Belgium, 1961.

———, and MCKIBBIN, D. D., An ion trap measurement of the ion concentration profile above the F_2 peak, *J. Geophys. Res.*, **66,** 1667–1671, 1961.

HARRIS, I., and PRIESTER, W., Theoretical models for the solar cycle variation of the upper atmosphere, *J. Geophys. Res.*, **67,** 4585–4591, 1962.

HOEGY, W. R., and BRACE, L. H., *Sci. Rept. JS-1, Univ. Mich. ORA Rept.* 03599–5–*S*, September 1961.

JACKSON, J. E., and BAUER, S. J., Rocket measurement of a daytime electron density profile up to 620 kilometers, *J. Geophys. Res.*, **66,** 3055–3057, 1961.

LINCOLN, J. V., Geomagnetic and solar data, *J. Geophys. Res.*, **66,** 311, 1961.

NICOLET, M., *Physics of the Upper Atmosphere,* chapter 2, Academic Press, New York, 1960.

RATCLIFFE, J. A., and WEEKES, K., *Physics of the Upper Atmosphere,* chapter 9, Academic Press, New York, 1960.

SPENCER, N. W., BRACE, L. H., and CARIGNAN, G. R., Electron temperature evidence for nonthermal equilibrium in the ionosphere, *J. Geophys. Res.*, **67,** 157–175, 1962.

ELECTRON DENSITY DISTRIBUTION IN THE UPPER F-REGION

S. CHANDRA*

Goddard Space Flight Center

A modified Chapman function with a variable scale height gradient has been found to be in good agreement with the electron density distribution obtained experimentally within the height range of about 100 km below the F2 peak to an altitude of about 700 km. The scale height distribution derived from this model is also consistent with the neutral gas scale height of the 1961 COSPAR International Reference Atmosphere.

INTRODUCTION

Although ionospheric investigations have been carried on extensively for about three decades, observational data were confined to the peak of the F-region and below, while the region above the peak remained unexplored due to the limitations of ground-based radio sounding techniques. Only recently, with the advent of rocket and radar backscatter techniques, has it been possible to explore the region above the F peak and to obtain some reliable observational data.

On the basis of their ion trap experiment from Wallops Island, Va., Hanson and McKibbin[1] concluded that the ion-density profile between 350 and 750 km corresponds to a hydrostatic distribution with a constant temperature of 1240°K assuming an ion mass of 16. Their measurement corresponded to 8:40 p.m., about 2 hours after sunset. Jackson and Bauer[2] obtained a daytime (3 p.m.) electron density profile between 220–620 km from the same station which supported the concept of a hydrostatic distribution above the F peak. Again assuming an ion mass of 16, they concluded the daytime temperature to be about 1640°K.

The purpose of this work is to examine electron density data in the light of the above findings and to estimate some of the atmospheric parameters. An analytical expression for the electron density distribution will be formulated in terms of the scale height which best fits the observational data and permits the determination of the scale height and other relevant parameters.

EMPIRICAL MODEL OF THE F-REGION

Wright[3] proposed a model of the ionosphere above the F2-maximum represented by a simple Chapman function with a constant scale height of 100 km. Yonezawa and Takahashi,[4] who developed an extensive theory of F2 region, showed that such a simplified version of the ionosphere was not consistent with experimental results in general. They attempted to fit an experimental profile obtained by Garriott[5] with a constant scale height gradient and found close agreement only up to 100 km above the F2 peak; the calculated and observed values were found to differ considerably in the higher region. Yonezawa and Takahashi suggested that a better fit with the experimental results might be obtained by assuming a variable scale height gradient.

With the availability a more reliable experimental data, it is now possible to examine these questions in more detail. In arriving at a model of the electron density profiles, from near the peak of the F2 layer to the region above, the following known features of the electron density distribution have been taken into account:

1. The region above 250 km is essentially isothermal and may be divided into three parts

* NAS—NASA Post-Doctoral Associate.

according to the relative concentrations of the ionic constituents. It is now generally believed that oxygen ions predominate up to an altitude of about 900 km, helium ions from 900 km to about 3000 km, and protons above 3000 km. The thicknesses of these regions vary with atmospheric temperature,[6] but the various ionic constituents are in thermodynamic equilibrium with the neutral atmosphere.

2. The electron distribution near the peak of the F layer can be approximated by a parabolic layer and that above the peak by an exponential layer varying as $\exp(-z/H')$, where H' is the scale height of electron ion gas. This region is usually called an isothermal diffusive equilibrium region. Since the predominant ionic constituent in this region is O^+ and the neutral atmosphere also consists predominantly of atomic oxygen, we may write, assuming thermodynamic equilibrium,

$$H' = 2H_O \tag{1}$$

where H_O is the scale height corresponding to atomic oxygen.

3. In the region near the F peak the scale height of the neutral atmosphere is height dependent because, even though the region is isothermal, the mean molecular mass changes with height. It approaches, H_O, the scale height of atomic oxygen at heights well above the peak.

We will begin with a Chapman function modified to include a variable scale height:

$$N = N_m \exp \frac{1}{2}\left[1 - \int_{h_m}^{h} \frac{dz}{H} - \exp\left(-\int_{h_m}^{h} \frac{dz}{H}\right)\right] \tag{2}$$

where N corresponds to electron density at height h and h_m is the height of the maximum electron density, N_m. The physical significance of the scale height h, will be discussed later. The following analytical form for H has been found to be compatible with the aforementioned features of the F region:

$$H = \frac{H_O\left[1 - \alpha \exp\left(-\frac{\alpha z'}{2}\right)\right]^2}{\left[1 - \alpha \exp\left(-\frac{\alpha z'}{2}\right) - \frac{\alpha^2 z'}{2} \exp\left(-\frac{\alpha z'}{2}\right)\right]}, \tag{3}$$

where

$$z' = \frac{z}{H_O} = \frac{h - h_m}{H_O},$$

$$\alpha = \frac{H_O - H_m}{H_O},$$

and H_m is the "scale height" at the peak ($z=0$). Equation 3 has been so formulated that with the increase of z', H approaches H_O and reduces to a constant scale height for $\alpha=0$. For $\alpha z' \ll 1$, equation 3 assumes a linear form. Substituting equation 3 in equation 1 gives

$$N = N_m \exp \frac{1}{2}\left\{1 - \frac{z'}{1 - \alpha \exp\left(-\frac{\alpha z'}{2}\right)} - \exp\left[\frac{z'}{1 - \alpha \exp\left(-\frac{\alpha z'}{2}\right)}\right]\right\}. \tag{4}$$

It is easy to verify that equation 4 reduces to the appropriate forms near the peak and well above the peak. In view of the different ion transition regions, equations 3 and 4 are assumed to be valid a few hundred kilometers above the peak where O^+ ions are predominant.

COMPARISON OF EXPERIMENTAL AND EMPIRICAL DATA

From Equation 4, N/N_m may be plotted as a function of z', for arbitrary values of α. This is shown in figure 1 for values of $\alpha=0$, 0.3, 0.4, 0.5, and 0.6. The curves corresponding to $\alpha=0$ is equivalent to the simple Chapman function. The parameter α may be interpreted as an index which measures the departure from the simple Chapman function.

In order to verify that equation 4 is in agreement with the measurements, we shall consider the electron density profile obtained by Jackson and Bauer[2] and the ion-density profile by Hanson and McKibbin,[1] shown in figures 2 and 3 respectively.

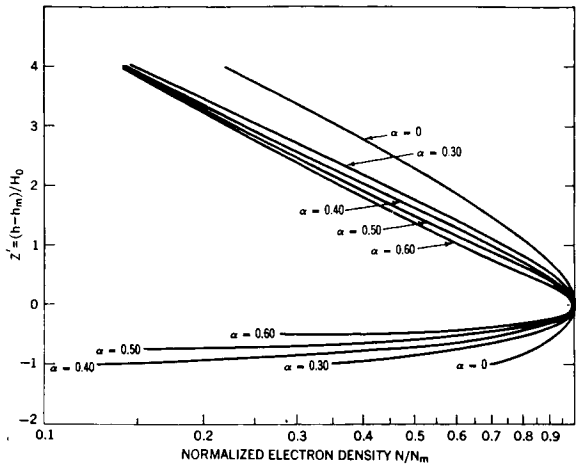


FIGURE 1.—Empirical models of electron density distribution.

Equation 3 is completely determined for a given set of values (N_m, h_m, α and H_O) which can be evaluated for any given experimental profile. The numerical values corresponding to the profiles shown are given in table 1. The computation was performed on the Goddard Space Flight Center IBM 7090 computer by a differential correction least-square fit.

In figures 2 and 3, equation 4 is compared with the experimental results. It is seen that the empirical formulation of electron density distribution is in excellent agreement with the experimental data. Note that the value of α is higher in the day than in the evening. Since α is an index

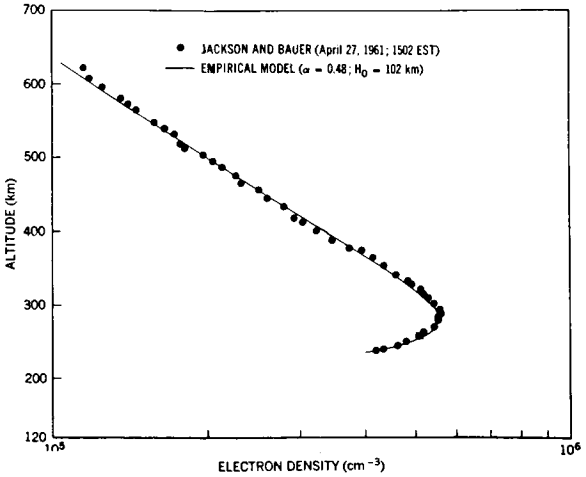


FIGURE 2.—Comparison of an experimental electron density distribution (reference 2) with the empirical model.

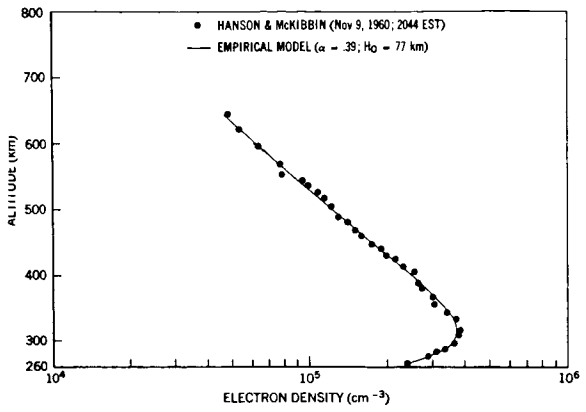


FIGURE 3.—Comparison of an experimental electron density distribution (reference 1) with the empirical model.

which measures the departure of equation 4 from the Chapman function, it may be concluded that the electron density distribution tends towards the Chapman function after sunset. This conclusion was also reached by Yonezawa, Martyn, Duncan, and Dungey[7–10] from purely theoretical considerations.

In addition to the two experimental profiles shown in figures 2 and 3, we have considered three sets of unpublished data[11,12] obtained during quiet sun conditions, to test the validity of equation 4.

In all three cases the Wallops Island Data (taken at 0947 and 2143 e.s.t., July 13, 1960 and 1817 e.s.t., June 24, 1961) were in excellent agreement with equation 4 between the height range of about 100 km below the peak to an altitude of about 700 km.

SCALE HEIGHT DISTRIBUTION

After the values of h_m, α and H_O are obtained for a given profile, equation 3 may be used to compute

TABLE 1.—*Numerical Values of the Parameters of Figures 2 and 3*

Profile	Time (e.s.t.)	N_m ($10^5 cm^{-3}$)	h_m (km)	h_m (km)	α
Electron density*----------	1500	5.55	280	102	0.48
Ion density†----	2040	3.73	313	77	0.39

*Reference 2.
†Reference 1.

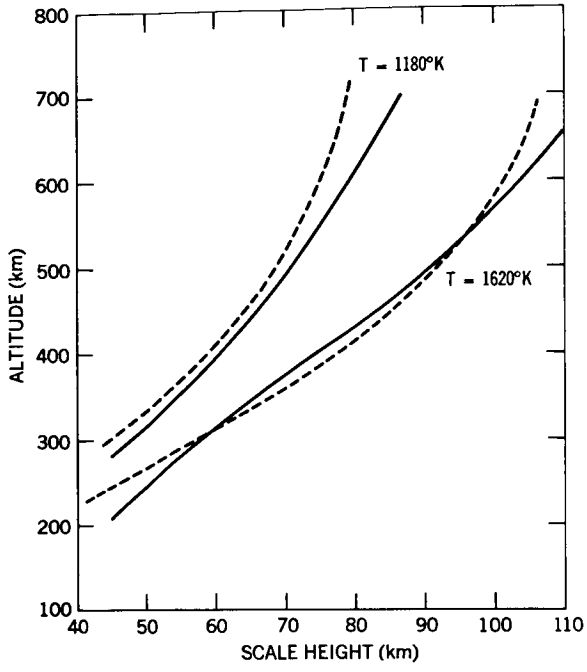


FIGURE 4.—Comparison of neutral scale heights of CIRA (solid curves) with the scale heights derived from electron-density data (dashed curves).

H as a function of height. The height variation of H is shown in figure 4 for the Jackson and Bauer, Hanson and McKibbin profiles and the neutral scale heights of the COSPAR International Reference Atmosphere.[13] Since H approaches the scale height of atomic oxygen in the region well above the peak, the corresponding isothermal temperature can be computed from the following relation

$$H = kT/mg, \tag{5}$$

where k is the Boltzmann constant, m the mean molecular mass, T the temperature and g the acceleration due to gravity. By taking $m = 16$ and the appropriate value of g at 600 km, the isothermal temperatures corresponding to the Jackson and Bauer and Hanson and McKibbin profiles are about 1620°K and 1180°K, respectively. These values are within the limits assigned by these authors to their temperature estimates. For the sake of proper comparison, the numerical values of the neutral scale heights of the reference atmosphere have been normalized to the temperatures in the isothermal region derived from the experimental data.

It is seen from figure 4, that the agreement between the neutral scale heights of CIRA and the scale heights derived from the electron density data is very good considering the model atmosphere uncertainties. It may therefore be reasonable to assume that the scale height derived from the electron density profile is actually the scale height of the neutral atmosphere. In the subsequent discussion we shall use H for the neutral scale height.

EQUATION OF CONTINUITY

The electron density and neutral scale height distribution obtained in the preceeding sections may now be examined in the light of the equation of continuity. In a horizontally stratified ionosphere, under the influence of the gravitational field only, the continuity equation may be written:

$$\frac{dN}{dt} = q - L - \frac{d}{dz}\left(NW_D\right), \tag{6}$$

where q and L are the rates of electron production and loss, respectively, and W_D is the upward ambipolar diffusion velocity of the electron ion gas, given by the following expression[14]:

$$W_D = -D\left[\frac{1}{N}\frac{dN}{dz} + \left(\frac{1}{T}\frac{dT}{dz} + \frac{1}{H'}\right)\right], \tag{7}$$

where H' is the scale height of the electron-ion gas, T' is the temperature and the diffusion coefficient

$$D = \frac{4.5 \times 10^{-17}}{n} T^{\frac{1}{2}} \sin^2 I, \tag{8}$$

where I is the magnetic dip angle and n is the density of the neutral atmosphere. For an isothermal region it can be easily shown that:

$$n = n_m \exp\left(-\int \frac{dz}{H}\right), \tag{9}$$

where n_m is the number density at the height of the maximum. Equation 6, together with equations 1, 2, 7, 8, and 9, may be written in the following form:

$$\frac{dN}{dt} = q - L + \frac{4.5 \times 10^{17}\, T^{\frac{1}{2}} \sin^2 T}{4H^2 n_m X}\left[X^2 - 2X\left(\frac{dH}{dz} - \frac{H}{2H_0} + 1\right) + \left(\frac{H}{H_0} + 2\frac{dH}{dz} - 1\right)\right], \tag{10}$$

TABLE 2.—*Numerical Estimates of the Production, Loss, and Diffusion Terms for Wallops Island*

Altitude (km)	$L(=\beta N)$ (cm^{-3} sec^{-1})	div NW_D (cm^{-3} sec^{-1})	$q = L + \text{div } NW_D$ (dm^{-3} sec^{-1})
255	124	18	142
280	84	37	121
306	47	36	83
331	25	29	54
357	14	22	36

where

$$X = \exp\left(-\int \frac{dz}{H}\right) .$$

The last term in equation 10 represents the effects of diffusion and may be computed for a given electron density profile and a given value of n_m.

Assuming that the loss term is an attachment type with a loss coefficient β given by Ratcliffe et al.[15]:

$$\beta = 10^{-4} \exp\left(-\int \frac{h - 300}{50}\right) \text{sec}^{-1}$$

between the height interval of 250–350 km, we can estimate the relative importance of the loss and diffusion terms. Further, for a quasi-equilibrium condition, it is also possible to estimate the rate of production. Thus, by using the values of N_m, h_m, α and T corresponding to figure 2 and assuming $n_m = 2 \times 10^9/\text{cm}^3$ at the peak ($h_m = 280$ km), the numerical estimates of the production, loss, and diffusion terms can be tabulated for Wallops Island, Va. ($I = 70°$). In table 2, the last column, which represents the production, is computed under the assumption of quasi-equilibrium.

The numerical values of q shown in the last column are in reasonable agreement with the estimates of the production rate of O^+ given by Watanabe and Hinteregger.[16]

CONCLUSION

An analytical form of the electron density distribution has been found which can be made to fit observational data, with high accuracy, within the height range of about 100 km below the F2 peak to an altitude of about 700 km. This technique may provide a useful tool for matching—in the region around the F2 maximum—the bottomside observations with the data from the Topside Sounder Satellite. In principle it is possible to use equation 4 to extrapolate the bottomside data of the quiet ionosphere up to an altitude of about 700 km.

The scale height distribution derived from the electron density data is consistent with the neutral scale heights of the COSPAR International Reference Atmosphere. By assuming that equation 4 is a solution of the equation of continuity and using appropriate values for diffusion and loss coefficients, the estimate of electron production rate is found to be in reasonable agreement with the values obtained from solar flux data by Watanabe and Hinteregger.[16] Consequently, it may be possible by further refinement of this method and definitive rocket flights containing simultaneous ionospheric, neutral atmosphere and solar radiation experiments, to use ionospheric data for studying the neutral atmosphere.

ACKNOWLEDGMENTS

The author wishes to thank Dr. S. J. Bauer for suggesting the problem and giving many useful suggestions. The author is also indebted to Messrs R. E. Bourdeau, J. E. Jackson, Dr. A. C. Aikin, and Mr. W. W. Berning.

REFERENCES

1. HANSON, W. B., and MCKIBBIN, D. D., "An Ion-Trap Measurement of the Ion Concentration Profile Above the F2-Peak," *J. Geophys. Res.*, **66**(6): 1667–1671, June 1961.
2. JACKSON, J. E., and BAUER, S. J., "Rocket Measurement of a Daytime Electron-Density Profile up to 620 Kilometers," *J. Geophys. Res.*, **66**(9): 3055–3057, September 1961.
3. WRIGHT, J. W., "A Model of the F Region Above h_{max} F2," *J. Geophys. Res.*, **65**(1): 185–191, January 1960.

4. YONEZAWA, T., and TAKAHASHI, H., "On the Electron and Ion Density Distributions From the Lower up to the Uppermost Part of the F Region," *J. Radio Res. Lab.*, **7**(32): 335–378, July 1960.
5. GARRIOTT, O. K., "Ionospheric Electron Content and Distribution Determined From Satellite Observations," in: *Space Research, Proc. of the 1st Internat. Space Science Symp., Nice, January* 1960, Amsterdam: North-Holland Publ. Co., 1960, 371–386.
6. BAUER, S. J., "On the Structure of the Topside Ionosphere," *J. Atmos. Sci.*, **19**(3): 276–278, May 1962.
7. YONEZAWA, T., "A New Theory of Formation of the F2 Layer," *J. Radio Res. Lab.* (*Japan*), **3**(11): 1–16, January 1956.
8. MARTYN, D. F., "Processes Controlling Ionization Distribution in the F2 Region of the Ionosphere," *Australian J. Phys.*, **9**(1): 161–165, 1956.
9. DUNCAN, R. A., "The Behaviour of a Chapman Layer in the Night F2 Region of the Ionosphere, Under the Influence of Gravity, Diffusion and Attachment," *Australian J. Phys.*, **9**(4): 436–439, December 1956.
10. DUNGEY, J. W., "The Effect of Ambipolar Diffusion in the Night-Time F Layer," *J. Atmos. Terrest. Phys.*, **9**(2/3): 90–172, 1956.
11. BERNING, W. W., ballistics research laboratory, technical report (to be published).
12. JACKSON, J. E., KNECHT, R. W., and RUSSELL, S., "First Results in NASA Topside Sounder Satellite Program," paper presented at the eighth annual meeting of the American Astronautical Society, Washington, January 1962 (to be published).
13. International Council of Scientific Unions, Committee on Space Research, Preparatory Group for an International Reference Atmosphere, "CIRA 1961: COSPAR International Reference Atmosphere, 1961;" report accepted at the COSPAR meeting, Florence, April 1961, comp. by H. Kallmann-Bijl, R. L. F. Boyd, et al., Amsterdam: North-Holland Publ. Co., 1961.
14. SHIMAZAKI, T., "A Theoretical Study of the Dynamical Structure of the Ionosphere," *J. Radio Research Lab.*, **6**(24): 109–241, March 1959.
15. RATCLIFFE, J. A., and SCHMERLING, E. R., et al., "The Rates of Production and Loss of Electrons in the F Region of the Ionosphere," *Philos. Trans. Roy. Soc.* London, **248A**, 621–642, 1956.
16. WATANABE, K., and HINTEREGGER, H. E., "Photoionization Rates in the E and F Regions," *J. Geophys. Res.*, **67**(3): 999–1006, March 1962.

PLASMA DIFFUSION IN THE IONOSPHERE

S. CHANDRA

Goddard Space Flight Center

Equations of motion appropriate to the conditions existing in the ionosphere are discussed with a view to examine the condition for ambipolar diffusion ($\mathbf{v}_e = \mathbf{v}_i$). It is shown that for quasi-equilibrium and isothermal conditions the required condition for ambipolar diffusion is given by curl $\mathbf{v} \times \mathbf{B} = 0$. It is further shown that the assumption of ambipolar diffusion along the field lines leads to the trivial situation of hydrostatic distribution of electron density independent of latitude.

These results are not in agreement with the generally accepted view that diffusion of the plasma along the direction of the magnetic field can account for many geophysical phenomena in the ionosphere. This disagreement is attributed to the fact that the assumption of field-aligned plasma diffusion puts a constraint on grad n_e which has not been taken into account by the previous workers. It is pointed out in the present paper that the solution of $\mathbf{v}_e$ and $\mathbf{v}_i$ in terms of the particle densities and temperatures are not possible without the knowledge of the electric field. The theoretical determination of the latter appears to be extremely complicated and it seems desirable to measure it experimentally.

INTRODUCTION

It now is generally accepted that diffusion plays an important role in controlling the distribution of ionization in the F-region of the ionosphere. The theory of diffusion appropriate to the conditions existing in the ionosphere was first proposed by Ferraro (1945) who treated the electron-ion gas as a single constituent and showed that the coefficient of diffusion of this gas is reduced by the ratio of $1:\sin^2 I$ in the presence of the magnetic field, I being the inclination (dip) of the earth's magnetic field. According to Ferraro, the vertical component of the velocity of diffusion v_z, for an isothermal condition is given by

$$v_z = -D \sin^2 I [n_e^{-1}(\partial n_e/\partial z) + H_1^{-1}] \qquad (1)$$

where D is the coefficient of diffusion, n_e the electron-density, H_1 the scale height of electron-ion gas, and z the altitude.

Based on equation (1) the diffusive equilibrium-distribution at places other than the magnetic equator is given by

$$n_e = n_{e0} \exp[-(z - z_0)/H_1] \qquad (2)$$

where n_{e0} is the electron-density at height z_0.

Equation (2) is in general accordance with the experimentally observed distribution well above the F_2-peak both at midlatitudes and above the equator even though according to equation (1), the vertical diffusion is inhibited at the geomagnetic equator.

Several modifications have been proposed to the original theory of Ferraro to take into account the temperature gradient and the effect of horizontal gardients but it has always been assumed or implied that the diffusion is essentially ambipolar ($\mathbf{v}_e = \mathbf{v}_i$), Johnson and Hulburt (1950), who treated the problem of plasma diffusion in the ionosphere in great detail showed that the electron-ion gas may diffuse together as a single constituent in the absence of the magnetic field, with a coefficient of diffusion and scale height twice that of the positive ions. In the presence of the magnetic field, however, diffusion is not ambipolar and is affected by the force exerted by the magnetic field on the electrical currents.

In view of the generally accepted conclusion that diffusion plays a significant role in controlling the charged-particle distribution in the upper ionosphere, it is important to examine if the condition for ambipolar diffusion actually exists in the

ionosphere where the effect of the magnetic field is significant. The purpose of this paper is to investigate this problem in detail and to specify the conditions which must be satisfied before the assumption of ambipolar diffusion can justifiably be used in the ionosphere.

EQUATION OF MOTION

In a multiple-component gas under the action of external forces, the equation of motion obeyed by each constituent may be written in the form proposed by Johnson (1951) and Schlüter (1951).

$$(\partial \mathbf{v}_s/\partial t)+\mathbf{v}_s\cdot(\partial \mathbf{v}_s/\partial \mathbf{r})=-\rho_s^{-1}(\partial/\partial \mathbf{r})\cdot\vec{\mathbf{p}}_s+\mathbf{F}_s+\sum_l \frac{m_l\nu_{sl}}{m_s+m_l}(\mathbf{v}_l-\mathbf{v}_s) \quad (3)$$

where the suffixes s and l stand for the type of the particles and the various terms in equation (3) may be defined as follows:

$\mathbf{v}_s$ = macroscopic velocity of the sth constituent
ρ_s = density
$\vec{\mathbf{p}}_s$ = pressure tensor; in general its ijth element is given by the following equation (Lamb, 1932)

$$(\mathbf{p}_s)_{ij}=[p_s\delta_{ij}+\tfrac{2}{3}\rho_s\eta_s(\partial v_{sk}/\partial x_k)\,\delta_{ij}]-\rho_s\eta_s[(\partial v_{is}/\partial x_j)+(\partial v_{sj}/\partial x_i)] \quad (4)$$

where η_s stands for the coefficient of kinetic viscosity.

$$\mathbf{F}_s=(e_s/m_s)\,(\mathbf{E}+\mathbf{v}_s\times\mathbf{B})+2\mathbf{v}_s\times\boldsymbol{\omega}-\Delta\Omega_{\text{tidal}}+\mathbf{g} \quad (5)$$

where $\mathbf{E}$, $\mathbf{B}$, e_s are, respectively, the electric field, the magnetic field and the charge, all expressed in MKS units, Ω_{tidal} is the tidal force due to the sun and the moon and $\mathbf{g}$ is the acceleration due to gravity. The terms $2\mathbf{v}_s\times\boldsymbol{\omega}$, known as the Coriolis acceleration appears because of the rotation of the terrestrial coordinate system with the angular frequency $\boldsymbol{\omega}$.

The last term in equation 3 represents the drag term. The symbols m_s and m_l are the masses of sth and lth kind of particles and ν_{sl} is the collision frequency of sth kind of particles with lth. The summation with respect to l is extended to cover all possible collision partners including $s=l$.

In discussing the problem of diffusion in the ionosphere we shall assume only three types of particles: electrons, ions, and neutral atoms. Further, to simplify our discussion we shall ignore the effect of viscosity, Coriolis, and tidal forces and consider the equation corresponding to quasi-equilibrium conditions. The neglect of viscosity removes the off-diagonal terms from the pressure term. Further, p may be treated as a scalar given by the equation of state

$$p_s=n_skT_s \quad (6)$$

where k is the Boltzmann constant and T_s the kinetic gas temperature.In a collision-dominated plasma the assumption of isotropicity of pressure is justified (Spitzer, 1962). Finally, we shall neglect the quadratic terms in $\mathbf{v}_s$ and its derivative, thereby, linearising all the equations. The approximation made so far are just the ones usually made in the study of the diffusion problem in the ionosphere. The equations of motion for neutrals, electrons, ions may now be written in the following form

$$n_nn_e\alpha_{en}(\mathbf{v}_n-\mathbf{v}_e)+n_nn_i\alpha_{in}(\mathbf{v}_n-\mathbf{v}_i)=-\nabla p_n+\rho_n\mathbf{g} \quad (7)$$

$$n_en_i\alpha_{ei}(\mathbf{v}_e-\mathbf{v}_i)+n_nn_e\alpha_{en}(\mathbf{v}_e-\mathbf{v}_n)=-\nabla p_e+\rho_e\mathbf{g}-en_e(\mathbf{E}+\mathbf{v}_e\times\mathbf{B}) \quad (8)$$

$$n_en_i\alpha_{ei}(\mathbf{v}_i-\mathbf{v}_e)+n_nn_i\alpha_{in}(\mathbf{v}_i-\mathbf{v}_n)=-\nabla p_i+\rho_i\mathbf{g}+n_ie(\mathbf{E}+\mathbf{v}_i\times\mathbf{B}) \quad (9)$$

where

$$\alpha_{ei}=[m_em_i\nu_{ei}/(m_e+m_i)n_i] \quad (10A)$$

$$\alpha_{en}=[m_em_n\nu_{en}/(m_e+m_n)n_n] \quad (10B)$$

$$\alpha_{in}=[m_im_n\nu_{in}/(m_i+m_n)n_n] \quad (10C)$$

The suffixes e, i, n in equations 7–10 stand for electron, ion and neutral molecule. For an isothermal condition, α_{ei}, etc. may be treated as constants.

The electric field $\mathbf{E}$, in general, is the sum of external and internal fields. Equations 7–9 should be supplemented by Maxwell's equations and equations of continuity. We may then write

$$\nabla\cdot\mathbf{E}=(n_i-n_e)\,(e/\epsilon_0) \quad (11)$$

where ϵ_0 is the free space permittivity

$$\nabla\times\mathbf{E}=0 \quad (12)$$

$$\nabla\cdot n_s\mathbf{v}_s=R_s \quad (13)$$

where R_s refers to the net volumetric rate of creation of sth kind of particles.

We may derive from equations 7–9, the following set of equations which will be useful in the subsequent discussion.

$$\mathbf{v}_n=\frac{n_e\alpha_{en}\mathbf{v}_e+n_i\alpha_{in}\mathbf{v}_i}{n_e\alpha_{en}+n_i\alpha_{in}}+\frac{\mathbf{A}_n}{n_n(n_e\alpha_{en}+n_i\alpha_{in})} \quad (14)$$

$$\mathbf{A}_e+\mathbf{A}_i+\mathbf{A}_n+e(n_i-n_e)\mathbf{E}+e(n_i\mathbf{v}_i-n_e\mathbf{v}_e)\times\mathbf{B}=0 \quad (15)$$

where

$$\mathbf{A}_{e,i,n}=-\nabla p_{e,i,n}+\rho_{e,i,n}\mathbf{g} \quad (16)$$

$$L(n_e\alpha_{en}+n_i\alpha_{in})(\mathbf{v}_e-\mathbf{v}_i)+e(\alpha_{en}\mathbf{v}_i+\alpha_{in}\mathbf{v}_e)\times\mathbf{B}$$
$$+e(\alpha_{en}+\alpha_{in})\mathbf{E}=\alpha_{in}(\mathbf{A}_e/n_e)-\alpha_{en}(\mathbf{A}_i/n_i) \quad (17)$$

where

$$L=\{\alpha_{ei}+[n_n\alpha_{en}\alpha_{in}/(n_e\alpha_{en}+n_i\alpha_{in})]\} \quad (18)$$

AMBIPOLAR DIFFUSION

In the following we shall investigate the condition under which the electron-ion plasma may diffuse together with a common velocity $\mathbf{v}$ (usually called ambipolar diffusion or plasma diffusion) such that

$$\mathbf{v}_e=\mathbf{v}_i=\mathbf{v} \quad (19)$$

substituting equation 19 in equation 17, we obtain

$$e(E+\mathbf{v}\times\mathbf{B})=\frac{\alpha_{in}}{\alpha_{en}+\alpha_{in}}\frac{\mathbf{A}_e}{n_e}-\frac{\alpha_{en}}{\alpha_{en}+\alpha_{in}}\frac{\mathbf{A}_i}{n_i} \quad (20)$$

It is clear that equation 20 must be satisfied if equation 19 is valid. The required condition therefore is given by

$$e\nabla\times\mathbf{v}\times\mathbf{B}=\nabla\times\left[\frac{\alpha_{in}\mathbf{A}_e/n_e}{\alpha_{en}+\alpha_{in}}-\frac{\alpha_{en}\mathbf{A}_i/n_i}{\alpha_{en}+\alpha_{in}}\right]=0 \quad (21)$$

since

$$\nabla\times\mathbf{E}=0$$

It is easy to verify that the R.H.S. of equation 21 is zero for an isothermal condition. Thus, the assumption of $\mathbf{v}_e=\mathbf{v}_i$ leads to the following condition.

$$\nabla\times\mathbf{v}\times\mathbf{B}=0 \quad (22)$$

This condition is always fulfilled in the absence of a magnetic field or when the motion is along the field lines. The last condition is generally assumed to be valid in the F-region and in the following we shall examine this case in detail.

DIFFUSION ALONG FIELD LINES

From equations 14 and 19

$$\mathbf{v}-\mathbf{v}_n=[-\mathbf{A}_n/(n_e\alpha_{en}+n_i\alpha_{in})n_n] \quad (23)$$

substituting equation 15 and equation 20 in equation 23 and assuming

$$m_e\ll m_i \quad \text{and} \quad (n_i-n_e)/n_{e,i}\ll 1,$$

we get

$$\mathbf{v}-\mathbf{v}_n\simeq-\frac{2kT}{n_n(\alpha_{en}+\alpha_{in})}\left(n_e^{-1}\nabla n_e-\frac{m_i\mathbf{g}}{2kT}\right) \quad (24)$$

In deriving equation 24, it is assumed that $T_e=T_i=T$. It is evident that equation 24 does not explicitly depend on the magnetic field. In the following, we assume for simplicity that $\mathbf{v}_n\ll\mathbf{v}$. This assumption may not be justifiable in general. However, it can be easily verified that equation 22 still holds even in this case. Equations 22 and 24 thus lead to the following equation

$$e\nabla\times\mathbf{v}\times\mathbf{B}$$
$$=\nabla\times\left[\frac{-2kT}{n_n(\alpha_{en}+\alpha_{in})}\left(n_e^{-1}\nabla n_e-\frac{m_i\mathbf{g}}{2kT}\right)\times\mathbf{B}\right]=0 \quad (25)$$

Equation 25 specifies the condition for ambipolar diffusion and must be solved to determine the required distribution. In the case of field-alined motion, i.e., when the plasma is diffusing along the field lines, equation 25 is clearly satisfied. We shall examine this case in the following.

In a spherical polar coordinate system coincident with the center of the earth, we may write

$$v_r=\frac{-2kT}{n_n(\alpha_{en}+\alpha_{in})}\left[n_e^{-1}\frac{\partial n_e}{\partial r}+\frac{m_i\mathbf{g}}{2kT}\right]$$

$$v_\theta=\frac{-2kT}{n_n(\alpha_{en}+\alpha_{in})}n_e^{-1}\frac{\partial n_e}{r\partial\theta} \quad (26)$$

$$v_\varphi=\frac{-2kT}{n_n(\alpha_{en}+\alpha_{in})}n_e^{-1}(r\sin\theta)^{-1}\frac{\partial n_e}{\partial\varphi}$$

where r is measured positive outward and θ and φ denote the geomagnetic colatitude and longitude. Further we may write

$$\mathbf{B} = -B(\mathbf{i}_r \sin I + \mathbf{i}_\theta \cos I) \tag{27}$$

where $\mathbf{i}_r$ and $\mathbf{i}_\theta$ are unit vectors along r and θ directions and I is the magnetic dip angle reckoned positive when the north-seeking pole of the needle points downward. The field-alined plasma diffusion case using equations 26 and 27 yields the following equation

$$n_e^{-1}(\partial n_e/\partial r) + (2H)^{-1} = n_e^{-1} r^{-1}(\partial n_e/\partial \theta) \tan I \tag{28}$$

where

$$H = kT/m_i \mathbf{g}$$

Equation 28 is a partial differential equation and we shall make use of it in solving the equation of continuity. Equation 13, putting $R_e = 0$, may be written

$$\nabla \cdot n_e \mathbf{v} = (1/r^2)(\partial/\partial r)(r^2 n_e v_r) + (r \sin\theta)^{-1}(\partial/\partial\theta)(n_e v_\theta \sin\theta) + (r\sin\theta)^{-1}(\partial/\partial\varphi)(n_\varphi v_\varphi) = 0 \tag{29}$$

Again from equations 15, 19, and 20 assuming $m_i \simeq m_n$ and neglecting n_e or n_i as compared to n_n we may write

$$n_n^{-1}(\partial n_n/\partial r) \simeq -m_n g/kT$$

$$n_n \simeq n_{n0} \exp -[(r - r_0)/H] \tag{30}$$

where n_{n0} refers to the neutral density at height r_0.

From equations 26, 28, 29, and 30 we get the following differential equation

$$[1 + \tfrac{1}{4}(\tan^2\theta)]\frac{\partial^2 n_e}{\partial r^2} + \frac{\partial n_e}{\partial r}\left[\frac{3}{2H} + \frac{3}{r} + \tfrac{1}{4}(\tan^2\theta)\left(\frac{3}{r} + H^{-1}\right)\right]$$

$$+ (n_e/2H)\{[H^{-1} + (3/r)] + \tfrac{1}{4}(\tan^2\theta)[(3/r) + (2H^{-1})]\} = 0 \tag{31}$$

In deriving equation 31, the dipole field approximation, i.e., $\tan I = 2\cot\theta$, has been assumed. The solution of equation (31) is given by

$$n_e(r, \theta) = A_1(\theta) \exp -(r/2H) + A_2(\theta) \exp -(r/2H) \int (r^3)^{-1} \exp -(2r/H)(4 + \tan^2\theta)\, dr \tag{32}$$

where $A_1(\theta)$ and $A_2(\theta)$ are two arbitrary functions of θ but are independent of r and must be evaluated to determine equation 32 uniquely. It is seen, however, that if equation 32 is substituted in equation 28 the only permissible values of $A_1(\theta)$ and $A_2(\theta)$ are when

$$A_1(\theta) = \text{const} = A_1$$

$$A_2(\theta) = 0$$

If we allow any other value of $A_2(\theta)$, the resulting differential equation of $A_2(\theta)$ has a solution which is not independent of r. This, however, is self-contradictory. We thus obtain

$$n_e = A_1 \exp(-r/2H) \tag{33}$$

A similar result may be obtained if equation 31 is written in terms of variable of θ instead of variable of r. In this case we obtain the following differential equation.

$$\frac{\partial^2 n_e}{\partial \theta^2} + \left[\frac{r}{H} - (1 + 4\cot^2\theta)\right]\frac{\cot\theta}{1 + 4\cot^2\theta}\frac{\partial n_e}{\partial \theta} = 0 \tag{34}$$

The solution of equation 34 may be written in the form

$$n_e(r, \theta) = B_1(r) + B_2(r) \int \tfrac{1}{4}(1 + 3\cos^2\theta)^{r/6H} \sin\theta\, d\theta \tag{35}$$

where $B_1(r)$ and $B_2(r)$ are arbitrary functions of r and must be evaluated from boundary conditions. Again substituting equation 35 in equation 28 we find, following the arguments given before,

that

$$B_2(r) = 0$$

$$N_e = B_1(r) = \text{const} \exp(-r/2H) \qquad (36)$$

which is equivalent to equation 33.

Equation 36 corresponds to the hydrostatic distribution of electron density with a scale height twice that of the neutral and is also obtained when $\mathbf{v}=0$.

As a result of the previous analysis it is clear that the assumption of ambipolar diffusion in the ionosphere requires that $\nabla\times\mathbf{v}\times\mathbf{B}=0$. In studying the effect of ambipolar diffusion on the charged particle distribution in the ionosphere, this point must be taken into account. In particular, the assumption of plasma diffusion along the field lines leads to the condition of hydrostatic distribution implying that $\mathbf{v}=0$, independent of geomagnetic latitude.

DISCUSSION

The results obtained in the preceding section are not in agreement with the findings of Kendall (1962), Lyon (1963), Rishbeth et al. (1963), and Goldberg and Schmerling (1963) even though all these authors have studied the case of plasma diffusion along the field lines. The disagreement arises due to the fact that they have not taken into account the additional constraint on ∇n_e given by equation 28 which automatically results from the assumption of field-alined plasma diffusion. However, in the opinion of this author this point must be taken into account. In view of the great geophysical importance of this problem we shall pursue this point a little further and investigate the limitations of solving this problem when $\mathbf{v}_e \neq \mathbf{v}_i$. For the conditions existing in the F-region where the gyro-frequency is much greater than the collision frequencies it is shown in the appendix that

$$\mathbf{v}_e = e(-\gamma_e+\delta_e\gamma_i)(\mathbf{E}\cdot\mathbf{h})\mathbf{h} - (\gamma_e+\delta_e\gamma_i)(n_e^{-1}\nabla p_e\cdot\mathbf{h})\mathbf{h} + (m_e\gamma_e+m_i\delta_e\gamma_i)(\mathbf{g}\cdot\mathbf{h})\mathbf{h} \qquad (37)$$

$$\mathbf{v}_i = e(\gamma_i-\delta_i\gamma_e)(E\cdot\mathbf{h})\mathbf{h} - (\gamma_i+\delta_i\gamma_e)(n_e^{-1}\nabla p_e\cdot\mathbf{h})\mathbf{h} + (m_i\gamma_i+m_e\delta_i\gamma_e)(\mathbf{g}\cdot\mathbf{h})\mathbf{h} \qquad (38)$$

where $\mathbf{h}$ is the unit vector along the field lines and the coefficient γ_e, γ_i etc. are defined by equation A-3 in the appendix. It is evident from equations 37 and 38 that both electron and ion velocities are along the magnetic field when their gyro-frequencies are much greater than their respective collision frequencies. However, it is not at all evident that the magnitudes of the two velocities are equal. It is not possible to make any further simplification of equations 37 and 38 since the mathematical problem leading to the solution for the electric field is extremely difficult. In order to avoid this difficulty one generally assumes $\mathbf{v}_e=\mathbf{v}_i=\mathbf{v}$. In this case the electric field can be easily eliminated from equations 37 and 38 leading to the well-known expression for the diffusion velocity.

$$\mathbf{v} = [\gamma_e\gamma_i/(\gamma_e+\gamma_i)]\{[(2/n_e)\nabla p_e + (m_e+m_i)\mathbf{g}]\cdot\mathbf{h}\}\mathbf{h} \qquad (39)$$

or in the component from

$$v_{er} = -D\left\{\left(n_e^{-1}\frac{\partial n_e}{\partial r}+T_e^{-1}\frac{\partial T_e}{\partial r}+(2H_i)^{-1}\right)\sin^2 I+\left(r^{-1}n_e^{-1}\frac{\partial n_e}{\partial\theta}+T_e^{-1}\frac{\partial T_e}{\partial\theta}\right)\cos I\sin I\right. \qquad (40)$$

$$v_{e\theta} = -D\left\{\left(n_e^{-1}\frac{\partial n_e}{\partial r}+T_e^{-1}\frac{\partial T_e}{\partial r}+(2H_i)^{-1}\right)\sin I\cos I+\left(r^{-1}n_e^{-1}\frac{\partial n_e}{\partial\theta}+T_e^{-1}\frac{\partial T_e}{\partial\theta}\right)\cos^2 I\right\}$$

where $D=2kT_e[\gamma_e\gamma_i/(\gamma_e+\gamma_i)]$ may be interpreted as diffusion coefficient.

Equation 40 has been the basis of studying the diffusion problem in the ionosphere. However, in the light of the discussion presented in this paper it is evident that under the assumption of $\mathbf{v}_e=\mathbf{v}_i$, the correct expression for $\mathbf{v}$ is given by equation 24 which does not explicitly depend on the magnetic field irrespective of the relative magnitude of gyro and collision frequencies. The

dependence on the magnetic field comes from equation 22 which is the required condition for $\mathbf{v}_e = \mathbf{v}_i$.

Unfortunately, there is no simple way of solving for $\mathbf{v}_e$ and $\mathbf{v}_i$ in terms of the particle densities and temperatures. In order to avoid the mathematical complexities, it may perhaps be desirable to devise experimental techniques for measuring the electric field. This will certainly be an important step in our understanding of the very intricate problem of diffusion in the ionosphere.

APPENDIX

If $\mathbf{v}_n \ll \mathbf{v}_e$ or $\mathbf{v}_i$, we may rewrite equations 8 and 9 in the following form.

$$\mathbf{v}_e + \lambda_e \mathbf{v}_e \times \mathbf{h} = \gamma_e [-e\mathbf{E} + (\mathbf{A}_e/n_e)] + \delta_e \mathbf{v}_i \quad \text{(A-1)}$$

$$\mathbf{v}_i + \lambda_i \mathbf{v}_i \times \mathbf{h} = \gamma_i [e\mathbf{E} + (\mathbf{A}_i/n_i)] + \delta_i \mathbf{v}_e \quad \text{(A-2)}$$

where $\mathbf{h}$ is a unit vector along the field lines and the coefficients γ_e, γ_i etc. are given by the following equations.

$$\begin{aligned} \gamma_e &= (n_i \alpha_{ei} + n_n \alpha_{en})^{-1} \\ \gamma_i &= (n_e \alpha_{ei} + n_n \alpha_{in})^{-1} \\ \lambda_e &= eB\gamma_e \\ \lambda_i &= -eB\gamma_i \\ \delta_e &= n_i \alpha_{ei} \gamma_e \\ \delta_i &= n_e \alpha_{ei} \gamma_i \end{aligned} \quad \text{(A-3)}$$

The coefficient λ_e and λ_i may be interpreted as the ratio of gyrofrequency and the effective collision frequencies of electrons and ions respectively. Further, δ_e and δ_i are the coupling terms between electron and ion motion through collision. Obviously, δ_e and δ_i are zero if $\mathbf{v}_e = \mathbf{v}_i$ since the terms containing α_{ei} are zero.

Equations A-1 and A-2 may be solved for $\mathbf{v}_e$ and $\mathbf{v}_i$. We thus obtain

$$\mathbf{v}_e = \frac{1-\lambda_e\lambda_i-\delta_e\delta_i}{(1-\lambda_e\lambda_i-\delta_e\delta_i)^2+(\lambda_e+\lambda_i)^2}\left\{\gamma_e \mathbf{G}_e\left[1+\frac{\lambda_i(\lambda_e+\lambda_i)}{1-\lambda_e\lambda_i-\delta_e\delta_i}\right]+\frac{\gamma_e(\mathbf{G}_e\cdot\mathbf{h})\mathbf{h}}{1-\delta_e\delta_i}\left[\frac{(\lambda_e+\lambda_i)(\lambda_e+\lambda_i\delta_e\delta_i)}{1-\lambda_e\lambda_i-\delta_e\delta_i}-\lambda_i\lambda_e\right]\right.$$

$$\left.+\gamma_e \mathbf{h}\times\mathbf{G}_e\left[\frac{\lambda_e+\lambda_i}{1-\lambda_e\lambda_i-\delta_e\delta_i}-\lambda_i\right]+\delta_e\gamma_i\mathbf{G}_i+\frac{\delta_e\gamma_i(\mathbf{G}_i\cdot\mathbf{h})\mathbf{h}}{1-\delta_e\delta_i}\left[\frac{(\lambda_e+\lambda_i)^2}{1-\lambda_e\lambda_i-\delta_e\delta_i}-\lambda_i\lambda_e\right]+\delta_e\gamma_i\mathbf{h}\times\mathbf{G}_i\left[\frac{\lambda_e+\lambda_i}{1-\lambda_e\lambda_i-\delta_e\delta_i}\right]\right\} \quad \text{(A-4)}$$

where

$$\begin{aligned} \mathbf{G}_e &= -e\mathbf{E} + (\mathbf{A}_e/n_e) \\ \mathbf{G}_i &= e\mathbf{E} + (\mathbf{A}_i/n_i) \end{aligned} \quad \text{(A-5)}$$

A similar expression can be written for $\mathbf{v}_i$ by interchanging the suffixes e and i.

At this stage it is appropriate to make a numerical estimate of the coefficients λ_e, γ_e, etc. in order to get a physical insight of equation A-4. For the conditions existing in the F-region, we adopt the following numerical values of the collision frequencies as given by Chapman (1956). These frequencies correspond to his model $h(T=1480°)$. Thus we may write

$$\begin{aligned} \nu_{ei} &= 268/\text{sec} \\ \nu_{en} &= 37.4/\text{sec} \\ \nu_{in} &= 1/\text{sec} \end{aligned}$$

Again

$$\omega_e = eB/m_e = 5.2\times10^6/\text{sec}$$

where $B = 0.3\times10^{-4}$ Weber

$$\omega_i = eB/m_i = 1.5\times10^2/\text{sec}$$

assuming $m_i = 19$ amu. Substituting these values in equation 39 and neglecting m_e as compared to m_i we obtain

$$\begin{aligned} \lambda_e &\simeq 1.7\times10^4 \\ \lambda_i &\simeq 2.8\times10^2 \\ \gamma_i &\simeq 0.6\times10^{26}\ \text{sec/kgm} \\ \gamma_e &\simeq 3.5\times10^{27}\ \text{sec/kgm} \\ \delta_e &\simeq .88 \\ \delta_i &\simeq 1.54\times10^{-2} \end{aligned}$$

From the above computations it is clear that λ_e

and $\lambda_i \gg 1$; $\delta_e \simeq 1$ and $\delta_i \ll 1$. With these approximations, equation A–4 can be written in the following from

$$\mathbf{v}_e = \gamma_e \left[\frac{\mathbf{G}_e}{\lambda_e^2} + (\mathbf{G}_e \cdot \mathbf{h})\mathbf{h} + \frac{\mathbf{h} \times \mathbf{G}_e}{\lambda_e} \right] + \delta_e \gamma_i \left[-\frac{\mathbf{G}_i}{\lambda_e \lambda_i} + (\mathbf{G}_i \cdot \mathbf{h})\mathbf{h} + \frac{\mathbf{h} \times \mathbf{G}_i}{\lambda_e \lambda_i} (\lambda_i^{-1} + \lambda_e^{-1}) \right] \quad \text{(A–6)}$$

A further approximation of equation A–6 results in the following equation.

$$\mathbf{v}_e \simeq \gamma_e (\mathbf{G}_e \cdot \mathbf{h})\mathbf{h} + \delta_e \gamma_i (\mathbf{G}_i \cdot \mathbf{h})\mathbf{h} \quad \text{(A–7)}$$

Following the same procedure for $\mathbf{v}_i$, we may write

$$\mathbf{v}_i \simeq \gamma_i (\mathbf{G}_i \cdot \mathbf{h})\mathbf{h} + \delta_i \gamma_e (\mathbf{G}_e \cdot \mathbf{h})\mathbf{h} \quad \text{(A–8)}$$

We may further write, assuming $n_e^{-1} \nabla p_e \simeq n_i^{-1} \nabla p_i$, and substituting for $\mathbf{G}_e$ and $\mathbf{G}_i$ from equation A–5

$$\mathbf{v}_e = e(-\gamma_e + \delta_e \gamma_i)(\mathbf{E} \cdot \mathbf{h})\mathbf{h} - (\gamma_e + \delta_e \gamma_i)(n_e^{-1} \nabla p_e \cdot \mathbf{h})\mathbf{h} + (m_e \gamma_e + m_i \delta_e \gamma_i)(\mathbf{g} \cdot \mathbf{h})\mathbf{h} \quad \text{(A–9)}$$

$$\mathbf{v}_i = e(\gamma_i - \delta_i \gamma_e)(E \cdot \mathbf{h})\mathbf{h} - (\gamma_i + \delta_i \gamma_e)(n_e^{-1} \nabla p_e \cdot \mathbf{h})\mathbf{h} + (m_i \gamma_i + m_e \delta_i \gamma_e)(\mathbf{g} \cdot \mathbf{h})\mathbf{h} \quad \text{(A–10)}$$

ACKNOWLEDGMENT

I am indebted to Dr. S. J. Bauer for giving many valuable suggestions throughout the course of this work. I also wish to thank Mr. W. Cahill, Dr. F. Mariani, and Dr. E. Maier for their critical comments.

This work was performed while the author held a NASA post doctoral research associateship from the National Academy of Sciences.

REFERENCES

Chapman, S., Nuovo Cimento, Suppl., **4,** 1385, 1956.

Ferraro, V. C. A., *Terr. Mag. and Atmos. Elect.*, **50,** 215, 1945.

Goldberg, R. A., and Schmerling, E. R., *J. Geophys. Res.*, **68,** 1927, 1963.

Johnson, M. H., and Hulburt, E. C., *Phys. Rev.*, **79,** 802, 1950.

———, Phys. Rev., **82,** 298, 1951.

Kendall, P. C., *J. Atmospheric Terrest. Phys.*, **24,** 805, 1962.

Lyon, A. J., *J. Geophys. Res.*, **68,** 2531, 1963.

Lamb, Hydrodynamics, Dover, 1932.

Rishbeth, H., Lyon, A. J., and Peart, Margret, *J. Geophys. Res.*, **68,** 2559, 1963.

Schlüter, A., *Z. Naturforshung*, **6a,** 73, 1951.

Spitzer, L., Physics of Fully Ionized Gases; Interscience Publishers, Inc., New York, 2d edition, 1962.

ANALYSIS OF TOPSIDE SOUNDER RECORDS

R. J. FITZENREITER AND L. J. BLUMLE

Goddard Space Flight Center

A large number of features can be seen on the ionograms obtained by the Canadian topside sounder satellite. The most useful feature for the calculation of electron density profile is the extraordinary trace; however, to achieve accurate results, the effect of the geomagnetic field must be carefully considered in the data reduction process. The method of analysis we selected assumes that small sections of the profile can be approximated by exponential laminations. This method requires fewer points than other first-order lamination techniques to achieve a given accuracy. It is usually assumed in the analysis that the received echoes correspond to vertical propagation. This assumption is not always valid, and it can in some cases lead to large errors. The sounder also excites the medium in the immediate vicinity of the satellite, giving rise to various plasma resonances. A graph is given which summarizes the local effects that are seen on the ionograms. This graph can be used for a rapid identification of these resonances.

INTRODUCTION

The success of the first topside sounder satellite, Alouette, launched on September 29, 1962, has provided a new means of studying the electron density distribution in the upper ionosphere (Warren, 1962; Petrie, 1963; Lockwood, 1963; Nelms, 1963; Muldrew, 1963a; Hagg, 1963). The principal equipment on Alouette is the ionospheric sounder, which sweeps from 0.5 to 11.5 Mc/s while the satellite moves about 80 km along its 1000-km circular, 80° inclination orbit; this sweep is repeated every 125 km along the orbital path. The sweep and orbital characteristics provide approximately one ionogram per degree of latitude for all except high latitudes. The orbital precession of nearly −8 min in time per day is such that, by combining southbound and northbound passes, we can obtain data at all local mean times every 3 months.

It is the purpose of this paper to discuss the interpretation of the data and to outline a technique for reducing Alouette data to electron density distributions with height.

CHARACTERISTICS OF TOPSIDE IONOGRAMS

Illustrated in figure 1 are two examples of typical Alouette ionograms observed at mid-latitudes and the corresponding N–h profiles. These ionograms differ from the conventional bottomside ionograms in that the virtual depth p', derived using the free-space wave velocity, is measured down from the satellite. The ordinary, extraordinary, and z-mode echoes are identified along with the plasma resonances and ground echoes.

The plasma resonances, first discussed by Lockwood (1963), occurs at the electron gyrofrequency and its harmonics and at frequencies where the phase refractive index is infinite or zero at the satellite. One additional phenomenon which is similar in appearance to a plasma resonance occurs on the ionograms when the refractive index is infinite for the vertical ray. The conditions for

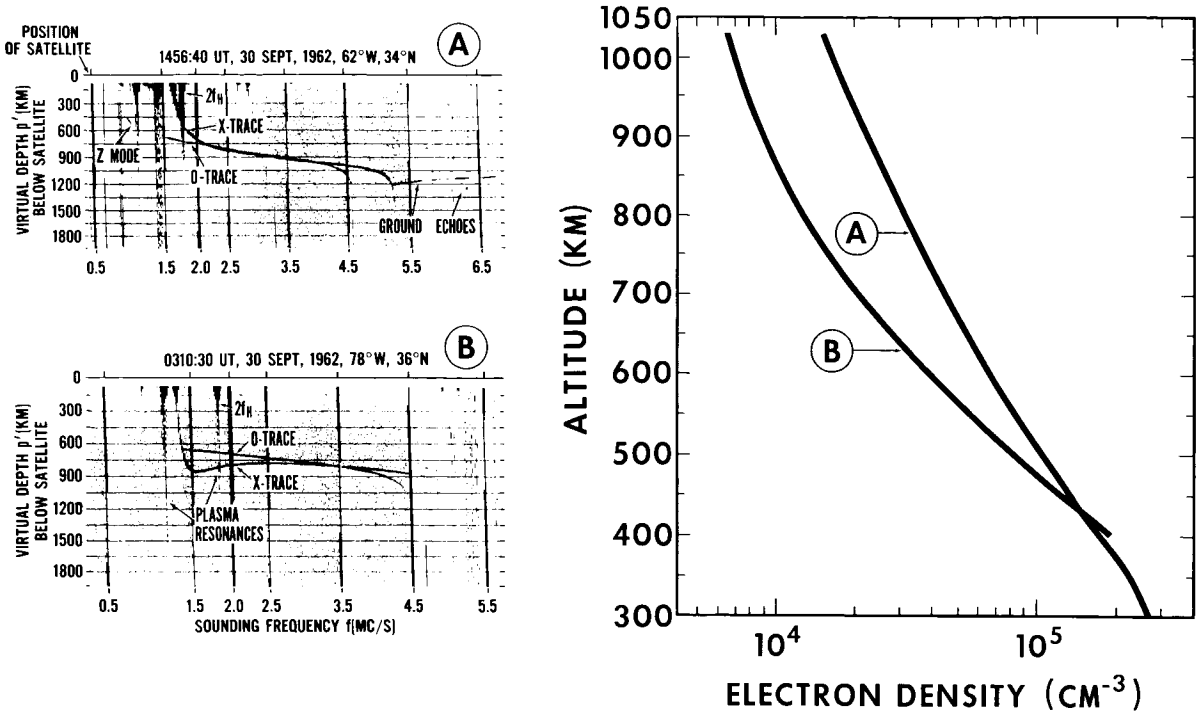


FIGURE 1.—Typical middle-latitude ionograms and corresponding true height profiles.

all these phenomena are given by the following formulas.

$$f_N^2 = f^2 \qquad X = 1 - f_N^2/f^2$$

$$f_N^2 = f^2 + f f_H \qquad X = 1 + Y = 1 + (f_H/f)$$

$$f_N^2 = f^2 - f f_H \qquad X = 1 - Y = 1 - (f_H/f)$$

$$f_N^2 = f^2 - f_H^2 \qquad X = 1 - Y^2 = 1 - (f_H^2/f^2)$$

$$f_N^2 = \frac{f^2 - f_H^2}{1 - (f_H^2 \cos^2\theta / f^2)} \qquad X = \frac{1 - Y^2}{1 - Y_L^2} = \frac{1 - (f_H^2/f^2)}{1 - (f_H^2 \cos^2\theta/f^2)}$$

where

f = wave frequency.
f_N = electron plasma frequency.
f_H = electron gyrofrequency.
θ = angle between the wave normal and the magnetic field.

The reflection conditions for the ordinary, z, and extraordinary modes (f_0, f_z, and f_x) are:

$$f_0 = 10^{-3}(80.6N)^{1/2}$$

$$f_z = 0.5\{-f_H + [f_0^2 + \tfrac{1}{4}(f_H^2)]^{1/2}\}$$

$$f_x = 0.5\{f_H + [f_0^2 + \tfrac{1}{4}(f_{H2})]^{1/2}\}$$

where N is the electron density per cubic centimeter, and all frequencies are in megacycles per second. These "resonance" conditions are represented graphically in figure 2, where the abscissa is the plasma frequency f_N, normalized with respect to the electron gyrofrequency f_H, and the ordinate is the normalized wave frequency f/f_H. With a graph of this type, the identification of resonances can be rapidly verified. The zeros of the refractive index are represented for the ordinary, extraordinary, and z mode by the lines labeled $X=1$, $X=1-Y$, and $X=1+Y$, respectively. The pole of the refractive index for the

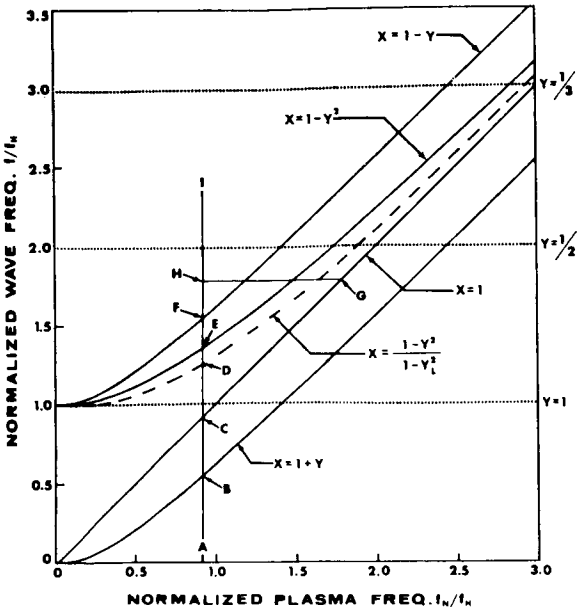


FIGURE 2.—Propagation and resonance conditions in a magnetoionic medium.

vertical ray occurs at $X=(1-Y^2)/(1-Y_L{}^2)$ and is illustrated (for $Y_L=Y\cos 45°$) by the dashed line. This condition can occur in the region bounded by the curves $X=1$, $Y=1$, and $X=1-Y^2$. At the magnetic equator it will coincide with the line $X=1-Y^2$, whereas at the magnetic dip pole it will occur along the $Y=1$ line if the normalized plasma frequency is less than unity and will coincide with the ordinary plasma resonance ($X=1$) for normalized plasma frequencies greater than unity. The line $X=1-Y^2$ represents the limit of z-mode propagation, and at this point the refractive index is infinite in the plane perpendicular to the magnetic field. The z mode can therefore only propagate in the region bounded by the curves $X=1+Y$ and $X=1-Y^2$, and the ordinary and extraordinary modes can propagate only in the regions above the curves $X=1$ and $X=1-Y$, respectively.

The plasma resonances observed at the electron gyrofrequency and its first two harmonics are represented by the dotted lines labeled $Y=1$, $\frac{1}{2}$, $\frac{1}{3}$. Others have shown that the $X=1$, $1-Y^2$, correspond to well-known plasma oscillations.

For a fixed electron density the propagation conditions encountered as the sounder sweeps from 0.5 to 11.5 Mc/s are represented by the vertical line AI on figure 2. As the sounder sweeps up in frequency the normalized wave frequency increases, but no propagation can take place until the point B on the curve $X=1+Y$ is reached, where propagation in the z mode can begin. This mode can propagate until $X=1-Y^2$ at point E, where z-mode propagation in any direction ceases. At point D, $X=(1-Y^2)/(1-Y_L{}^2)$ for the vertical ray (at 45° dip), the group retardation becomes very large as the energy propagates down to the reflection point $X=1+Y$ and returns. The process of propagating vertically to a reflection point and back can be represented by a horizontal line such as HG (for the ordinary ray) if we neglect the variation of gyrofrequency with altitude and further assume that plasma frequency increases in the direction of propagation. Propagation in the ordinary and extraordinary modes begins when $X\leq 1$ and $X\leq 1-Y$ (points C and F), respectively.

The gyroresonances can occur when the sounder frequency is an integral multiple of the electron gyrofrequency; i.e., $Y=1$, $\frac{1}{2}$, $\frac{1}{3}$, etc. Alouette observations of the gyrofrequency resonances, phenomena that are well known in plasma physics research (Wharton, 1959), have been qualitatively explained by Lockwood (1963). The observed "resonance" at $X=(1-Y^2)/(1-Y_L{}^2)$ for the vertical ray can be simply explained by the large

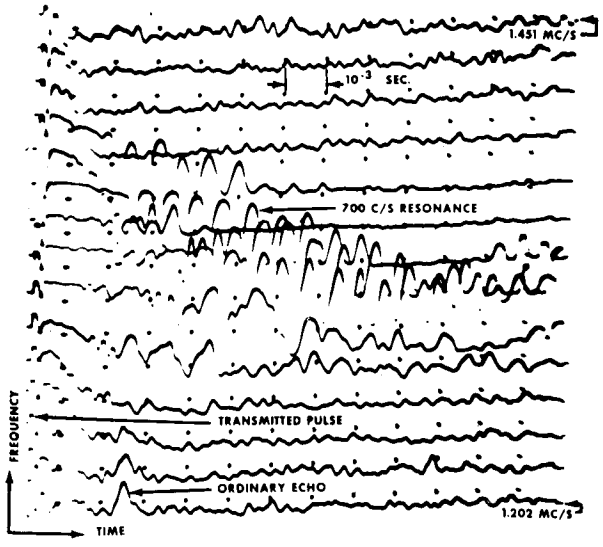


FIGURE 3.—Low-frequency oscillation observed in a plasma resonance.

group retardation encountered by the z mode as it propagates to the reflection point and returns.

Such a resonance and its fine structure are shown in figure 3, which is a photograph of 15 successive soundings separated by about 16 kc/s in wave frequency. This figure is a plot of signal amplitude (ordinate) as a function of time (abscissa) as the sounder sweeps from 1.202 to 1.451 Mc/s. One-millisecond timing marks enable the spectral components of this plasma resonance to be easily determined. The principal component of this resonance, 700 cps, does not change as the wave frequency increases 16 kc/s per pulse. The hydrogen ion gyrofrequency for this geomagnetic location is within 5% of the measured principal spectral component. For this location (133.5°E, 16.5°S) and time (1623 UT, March 11, 1963), hydrogen is believed to be the predominant ion at the satellite altitude (Bauer, 1962). On the basis of these considerations, the observed periodicity may indeed be the hydrogen ion gyroresonance. This phenomenon is seen most frequently in geomagnetic locations where the magnitude of the field is largest, and on the conventional ionogram display it appears as a broken vertical line. The possibility that this oscillation is introduced by the satellite receiver is being investigated.

As can be seen from figure 1, the z-mode trace does not appear on every ionogram, and the ordinary mode echoes usually are not seen at frequencies much lower than the extraordinary exit frequency ($X=1-Y$). J. W. King (private communication) has observed some ionograms near the magnetic equator on which the ordinary mode echoes extend back to the ordinary exit frequency, $X=1$.

The echoes observed above 4.5 Mc/s (f_0F_2) on the ionogram shown in figure 1A are returns from the ground; similar echoes have also been observed from sporadic E ionization on some ionograms. Spread F is observed from the topside and has been discussed by Petrie (1963) and Knecht and Van Zandt (1963). As will be indicated later, oblique echoes have also been seen.

REDUCTION OF p'—f RECORDS TO $N(h)$ PROFILES

The initial step in the analysis of Alouette data is the conversion of sounding records to electron density distributions with altitude from which various ionospheric parameters can be inferred. The relation between the measured virtual depth of reflection p' from the satellite and the true depth of reflection p is

$$p' = \int_0^p \mu'(f, N, B, \theta)\, dp \qquad (1)$$

The group refractive index, μ', given by the ratio of the free-space wave velocity to the group velocity, is a function of the wave frequency f, the electron density N, the magnitude of the magnetic field B, and the angle θ between the wave normal and the direction of B. For vertical propagation, θ is the complement of the magnetic dip angle. Although the group velocity V_g is a known function of electron density (Shinn and Whale, 1952), the determination of the depth p at which reflection occurs; i.e., the point at which $V_g=0$, requires advance knowledge of the electron density distribution $N(p)$. The heart of the p'—f reduction problem therefore lies in the inversion of the integral in (1) that is necessary to determine p.

The problem must be approached by making assumptions about the electron density distribution, thereby approximating the $N(p)$ profile by a model. Various methods have been developed for the reduction of bottomside ionograms, and with appropriate modifications they can be applied to topside ionogram reduction. These methods fall roughly into two categories corresponding to the approximation made, viz, polynomial and lamination techniques. The polynomial technique developed independently by Unz (1961), Titheridge (1961), and Knecht et al. (1962) assumes that the entire profile can be approximated by a single polynomial in plasma frequency f_N. A modification of the single polynomial technique (Titheridge, 1961) approximates the profile by a number of overlapping polynomials. Thomas et al. (1963) have applied the single polynomial technique to topside reduction for the special case in which the geomagnetic field is neglected. The lamination method originally developed by Budden (1955) assumes that the profile can be approximated by a number of slabs or laminations, the height increments of which are a function of electron density. Budden assumed the height increments to be linear in plasma frequency, f_N (where

$f_N = (8.98 \times 10^{-3}) N^{1/2}$ with f_N in megacycles per second and N in electrons per cubic centimeter); this technique has been used extensively by Schmerling (1957) and others. Jackson (1956) has developed a reduction method similar in principle with height increments linear in electron density. These lamination methods should properly be called first-order lamination methods, since the height increments are linear in plasma frequency or in electron density. Other lamination techniques developed by Paul (1960) and Doupnik (1963) assume height increments to be a parabolic function of electron density.

In applying a lamination method to the reduction of topside ionograms, it is logical to assume the height increments to be a function of the natural logarithm of electron density, $\log(N)$, since theoretical considerations (Bauer, 1962) and rocket experiments (Jackson and Bauer, 1961; Bauer and Jackson, 1962) indicate that the topside electron density distribution is essentially exponential in character. A similar exponential lamination technique has been proposed by King (1960) for the analysis of bottomside ionograms. The following sections outline the first-order linear-in-$\log(N)$ lamination method and its application to topside sounder ionogram reduction.

EXPONENTIAL LAMINATION METHOD

Lamination methods of $p'-f$ reduction enable the integral in (1) to be approximated by a summation, i.e.,

$$p'_i = \int_0^{p_i} \mu'(f_i, N, B, \theta)\, dp$$

$$\approx \sum_{j=1}^{i} \bar{\mu}'_{ij}(p_j - p_{j-1}) \qquad (2)$$

where p_i and p'_i are the true depth of reflection and virtual depth, respectively, for a wave with frequency f_i reflected at an electron density N_i, and $p_0 = p'_0 = 0$. $\bar{\mu}'_{ij}$ is the group refractive index for the frequency f_i as the wave passes through the slab defined by the true depths p_{j-1} and p_j. For example, $\bar{\mu}'_{53}$ represents the average value of μ' in the third slab for the frequency which is reflected at the bottom of the fifth slab. The expression for $\bar{\mu}'_{ij}$ is

$$\bar{\mu}'_{ij} \quad \frac{1}{p_j - p_{j-1}} \int_{p_{j-1}}^{p_j} \mu'(f_i, N, B, \theta)\, dp \qquad (3)$$

The calculation of $\bar{\mu}'_{ij}$ depends on the electron density distribution assumed within the slab. The linear-in-$\log(N)$ technique described here assumes that $N(p)$ is a simple exponential within each slab, i.e.,

$$N = N_{j-1} \exp[k_j(p - p_{j-1})] \qquad (4)$$

where N_{j-1} is the electron density at p_{j-1}, the distance from the satellite to the top of the slab, and $(k_j)^{-1} = (p_j - p_{j-1})/[\log(N_j) - \log(N_{j-1})]$ is the assumed constant electron-ion scale height for the slab. Solving (4) for p, differentiating with respect to $\log(N)$, and substituting the result into (3) yields

$$\bar{\mu}'_{ij} = [\log(N_j/N_{j-1})]^{-1} \cdot \int_{\log(N_{j-1})}^{\log(N_j)} \mu'(f_i, N, B, \theta)\, dN/N \qquad (5)$$

APPLICATION OF THE EXPONENTIAL LAMINATION METHOD

The principal difference between the reduction of topside sounder ionograms and conventional bottomside ionograms is that in the topside region the plasma frequency is nonzero over the entire ray path. A further complication is that the magnitude of the geomagnetic field varies appreciably over the ray path. For reduction of the ordinary trace, the field variation can be ignored for a first-order solution, but for reduction of the extraordinary trace the magnetic field is a very important parameter.

The electron density at the satellite is obtained from

$$N_0 = (1.2404 \times 10^4) f_{0v}^2 \qquad (6)$$

for the ordinary trace ($X = 1$) and

$$N_0 = (1.2404 \times 10^4)(f_{xv}^2 - f_{xv} f_{Hv}) \qquad (7)$$

for the extraordinary trace ($X = 1 - Y$). The terms f_{0v} and f_{xv} are the wave frequencies in megacycles per second for the ordinary and extraordinary modes, respectively, at which the virtual depth is zero, and f_{Hv} is the electron gyrofrequency at

the vehicle. The local gyrofrequency can be obtained from an inverse cube extrapolation of the ground value, from gyrofrequency plasma resonances, or by means of a polynomial fit to the geomagnetic field.

Since ionograms do not in general show an ordinary trace for wave frequencies much less than that at which propagation in the extraordinary mode begins, the electron density at the satellite can therefore only be determined by (7).

To begin the calculation of the profile, we compute the average group refractive index $\bar{\mu}'_{11}$ from (5) for the first lamination which is bounded by the electron density N_0 at the satellite and the electron density N_1 corresponding to the frequency of the first point scaled from the given $p'-f$ record. From (2) for $i=1$, the true depth is given by $p_1=p'_1\bar{\mu}'_{11}$. The true altitude h_1 is found by subtracting p_1 from the satellite altitude h_v. To calculate the point on the profile corresponding to the reflection of a wave frequency f_i, the contribution to the retardation of f_i in the last lamination before reflection, $\bar{\mu}'_{ii}(p_i-p_{i-1})$, is found by subtracting from p'_i the retardation due to that part of the profile already determined; from this result and the average group refractive index for the last lamination, p_i and h_i are then found by the following two equations:

$$p_i=p_{i-1}+\left[p'_i-\sum_{j=1}^{i-1}\bar{\mu}'_{ij}(p_j-p_{j-1})/\bar{\mu}'_{ii}\right] \quad (8)$$

$$h_i=h_v-p_i \quad (9)$$

This process is continued until the entire profile is determined.

MAGNETIC FIELD EFFECTS

Although the dip angle for a given geographic location is nearly constant within the altitude range of topside soundings, the variation of the electron gyrofrequency can be as much as 25% over the range 250 to 1000 km. Therefore the accurate determination of true height profiles requires allowing for the altitude variation of the magnitude of the field in the calculation of the integrals in (5). This variation can be taken into account by evaluating f_H at the altitude corresponding to the bottom of each lamination and assuming it to be constant throughout the lamination. This value of f_H is used in computing $\bar{\mu}'$ and, when the propagation is in the extraordinary mode, f_H is also required to compute the upper limit of the integral of μ' in (5). A slight complication arises in computing the retardation for the last lamination, since the altitude of reflection (and hence f_H) is not known initially. This difficulty is overcome in the following manner: The average group refractive index $\bar{\mu}'_{ii}$ is calculated using f_H evaluated at the altitude corresponding to the top of the layer, h_{i-1}, and from (8) and (9) a value for h_i is obtained. The value of $\bar{\mu}'_{ii}$ is then recomputed using f_H evaluated at this initial h_i, and a new value of h_i is determined. Continuing this iterative process will not improve the degree of accuracy appreciably because of the approximation that f_H is a constant throughout the slab. Doupnik (private communication) has shown that by using only the extraordinary virtual depth for each lamination it is possible to improve the approximation by calculating the field as a function of altitude within the lamination, since the slope of the electron density profile for each lamination is the only unknown.

In most cases, the field compensation scheme described above is sufficiently accurate, i.e., the uncertainties in the true height calculations due to the field dependence are less than the uncertainties inherent in the reduction method. It has been found, however, that at locations where the extraordinary exit frequency and the local electron gyrofrequency are such that the field parameter $Y=f_H/f$ is greater than about 0.7, the uncertainty in the first few true heights of reflection below the satellite (+20 km) is greater than the combined uncertainty due to scaling, computational, and satellite position errors ($\pm$10 km). This uncertainty is due to the extreme sensitivity of the index μ' and of the electron density calculations to slight deviations in Y when Y approaches unity. As the sounding frequency increases, Y decreases, and the uncertainties in the remaining true height calculations due to the field dependence become small.

Reduction of the ordinary ray data is much simpler, since the electron density at the bottom of the lamination is known, and therefore the height variation of Y within the lamination can

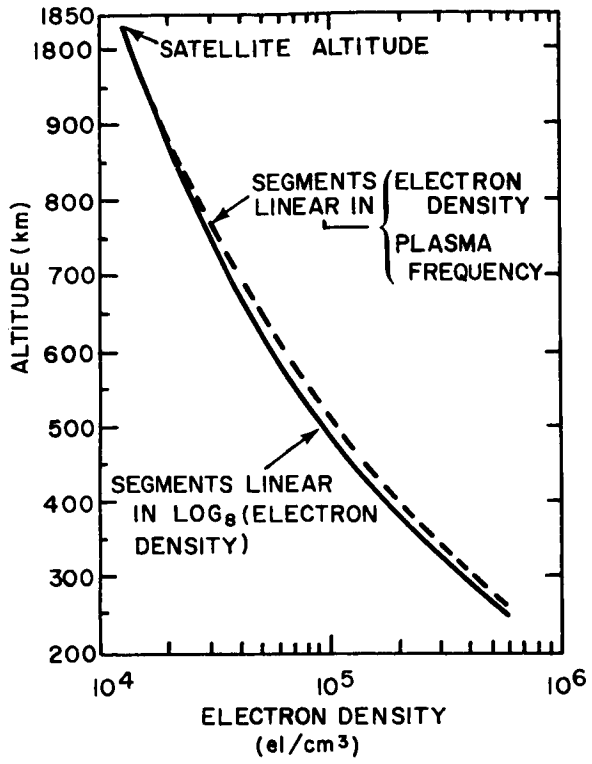


FIGURE 4.—Comparison of first-order lamination methods as applied to topside ionogram reduction.

be included more accurately by an iterative procedure. For the ordinary ray, the Y dependence of $d\mu'/dY$ for $\mu'>2$ is large for two ranges of Y, $\frac{1}{8}\lesssim Y\lesssim\frac{1}{2}$ and $Y>2$. The range $Y>2$ is not encountered with the current topside sounder, and $Y\cong\frac{1}{2}$ is seldom encountered in the altitude range where $\mu'>2$. Including the height variation of Y within each lamination does, however, reduce the error in a true height profile (for which $Y=\frac{1}{2}$ at the satellite) by nearly 3 km over the entire profile.

COMPARISON OF LAMINATION METHODS APPLIED TO TOPSIDE IONOGRAMS

When applying a lamination method to ionogram reduction, the error introduced by the departure of the assumed from the actual electron density distribution within each lamination can be determined by the number of laminations required to define the profile accurately. A comparison of the three first-order lamination reduction techniques applied to the extraordinary traces of the same ionogram is shown in figure 4. The solid curve is the profile obtained by the linear-in-log(N) technique, while a single dashed curve shows the profiles obtained by the linear-in-N and linear-in-f_N techniques, since these last two agree with each other to within 4 km. The maximum difference in altitude between the two curves is about 20 km. Sixteen laminations were used in all three techniques. Doubling the number of laminations in all three techniques yields profiles which are identical (within 3 km) to the profile represented by the solid curve. Therefore, to calculate profiles of the same accuracy, fewer laminations are required by the linear-in-log(N) technique than by the linear-in-N or linear-in-f_N techniques. If a height variation of higher order in log(N) were assumed, it is expected that even fewer laminations would be required.

N(h) PROFILES

All methods of reducing ionograms to $N-h$ profiles assume that the wave normal is in the vertical direction. There is, however, evidence on some ionograms of field-aligned propagation. Muldrew (1963b) has discussed observations of trapping in field-aligned ducts near the magnetic equator. Figure 5 is an observed topside ionogram and two electron density distributions derived therefrom. Profile A, obtained on the assumption that the propagation path was entirely vertical, is obviously in error because it yields an F_2 maximum of 10^5 el/cm^3 at 100 km. Profile B, obtained on the assumption of field-alined propagation, results in an F_2 maximum at 251 km. Although there is no direct evidence on this ionogram that the propagation is field alined, this conclusion, based on the height of the F_2 maximum, is the only reasonable estimate of the actual conditions. It should be emphasized again that the overall error in any data point due to scaling, computations, etc., is less than 10 km.

It is well known that a cusp will appear on a virtual height record whenever the slope of the electron density distribution is large and rapidly changing. The depth of this cusp depends principally on the slope, and the shape is a measure of the rate of change of the slope. The large cusp in figure 5 at 1.65 Mc/s is produced by the change

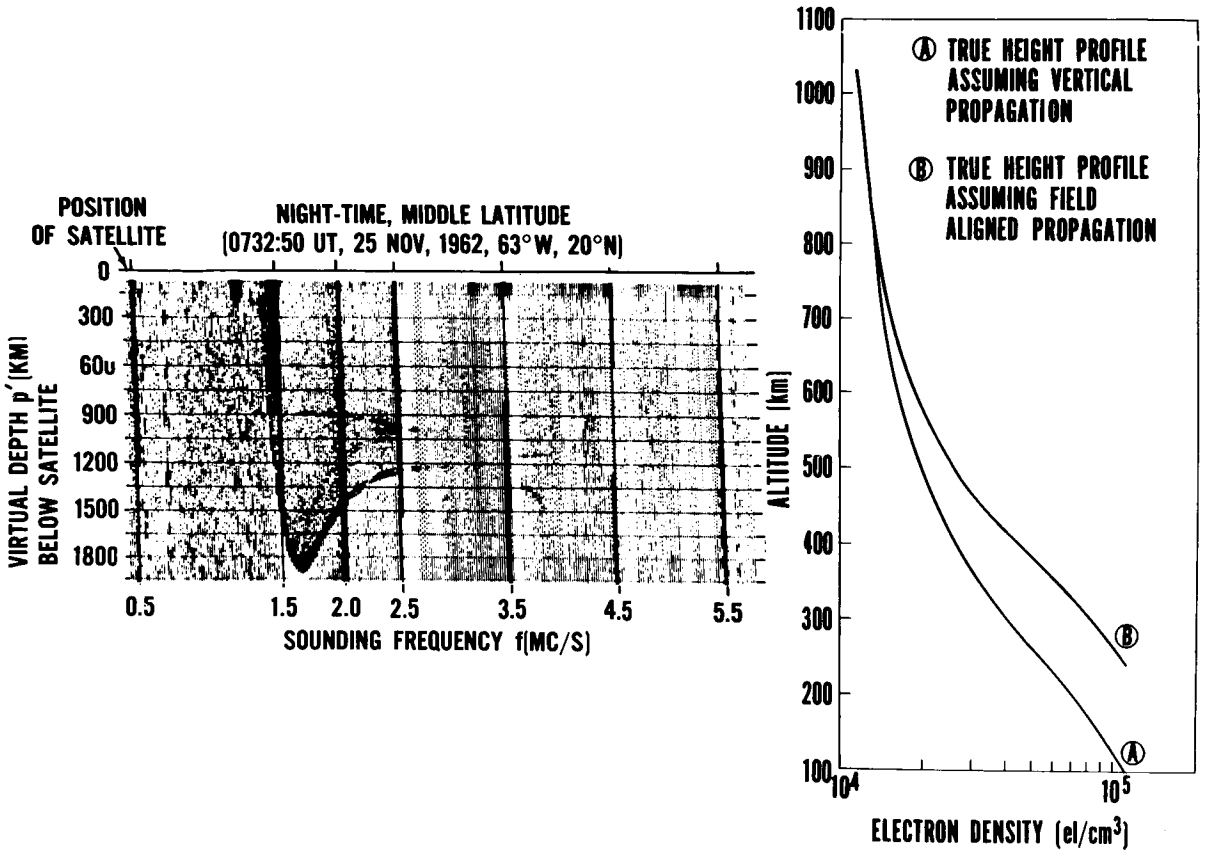


FIGURE 5.—One example of appreciable deviation from vertical propagation.

in scale height between 600 and 800 km. This cusp does not arise from the large values of the group refractive index that occur at $Y=1$, since the largest value of Y for this sounding is 0.7. If the electron density distribution and electron gyrofrequency were such that $0.9<Y<1.0$, a sharp cusp would be observed in the extraordinary trace due to the variation of Y; however, this phenomenon has not yet been observed, since the electron densities observed at 1000 km are generally greater than 10^3 el/cm^3.

SUMMARY AND CONCLUSION

Topside ionograms exhibit local phenomena (plasma resonances), echoes from the topside ionosphere (ordinary, extraordinary, and z propagation modes), and echoes from below the F_2 maximum (ground returns and sporadic E reflections). The range of extraordinary echoes extends rather consistently from the immediate vicinity of the satellite to the reflection point, and it is thus the most useful for calculating electron density profiles. Lamination methods of reducing $p'-f$ records to $N-h$ profiles have been discussed and, in particular, a method using logarithmic laminations has been developed and applied to the topside ionosphere. This method yields greater accuracy than other first-order lamination techniques for a small number of sample points. The height variation of the geomagnetic field is important in (1) the computation of the average group refractive index in each lamination and (2) the calculation of the electron density at the boundaries of each lamination in the case of the extraordinary ray.

A convenient means of identifying the plasma resonances observed with a swept frequency topside sounder has been presented. In one of the plasma resonances a spectral component has been

observed that appears to correspond to the H^+ gyrofrequency.

Some topside ionograms, reduced under the assumption of vertical propagation, yield profiles with a value of the F_2 peak that is obviously too low. It is suggested that in these cases field-aligned propagation is the source of this error.

ACKNOWLEDGMENTS

The continuous operation of the swept frequency topside sounder satellite launched more than a year ago is a tribute itself to the excellent work of the Defence Research Telecommunications Establishment. Many helpful discussions with J. E. Jackson and S. J. Bauer of the Goddard Space Flight Center are hereby acknowledged. We are indebted to J. Doupnik for the group refraction index computer subroutine, whose accuracy was verified by comparing with Becker's (1960) values.

REFERENCES

Bauer, S. J., On the structure of the topside ionosphere, *J. Atmospheric Sci.*, **19,** 276, 1962.

———, and Jackson, J. E., Rocket measurement of the electron density distribution in the topside ionosphere, *J. Geophys. Res.*, **67,** 1675, 1962.

Becker, W., Tables of ordinary and extraordinary refractive indices, group refractive indices and $h_{0,x'}(f)$-curves for standard ionospheric layer models, *Max-Planck Inst. Aeronomy*, No. 4, 1960.

Budden, K. G., *Cambridge Conference on Ionospheric Physics*, p. 332, Physical Society, London, 1955.

Doupnik, J. R., A flexible method of determining the electron density in the ionosphere, *Penna. State Univ. Sci. Rept.*, **190,** 1963.

Hagg, E. L., A preliminary study of the electron density at 1000 kilometers, *Can. J. Phys.*, **41,** 195–199, 1963.

Jackson, J. E. A new method for obtaining electron-density profiles from *p–f* records, *J. Geophys. Res.*, **61,** 107, 1956.

———, and Bauer, S. J., Rocket measurements of a daytime electron density profile up to 620 kilometers, *J. Geophys. Res.*, **66,** 3055, 1961.

King, G. A. M., Use of logarithmic frequency spacing in ionogram analysis, *J. Res. NBS*, **64**D, 501, 1960.

Knecht, R. W., and Van Zandt, T. E., Some early results from the ionospheric topside sounder satellite, *Nature*, **197,** 641, 1963.

———, ———, and Watts, J. M., Electron density profiles in the ionosphere and exosphere, *NATO Conf. Ser.*, **2,** 246, 1962.

Lockwood, G. E. K., Plasma and cyclotron spike phenomena observed in topside ionograms, *Can. J. Phys.*, **41,** 190–194, 1963.

Muldrew, D. B., The relationship of F-layer critical frequencies to the intensity of the outer Van Allen belt, *Can. J. Phys.*, **41,** 199–202, 1963a.

———, Radio propagation along magnetic field-aligned sheets of ionization observed by the Alouette topside sounder, *J. Geophys. Res.*, **68,** 5355–5370, 1963b.

Nelms, G. L., Scale heights of the upper ionosphere from topside soundings, *Can. J. Phys.*, **41,** 202–206, 1963.

Paul, A. K., Aktive Hochfrequenzspektrometer fur Ionospharische Echolotung, *A.E.U.*, **14,** 468, 1960.

Petrie, L. E., Topside spread echoes, *Can. J. Phys.*, **41,** 194–195, 1963.

Ratcliffe, J. A., *The Magnetoionic Theory and its Applications to the Ionosphere*, Cambridge University Press, London, 1959.

Schmerling, E. R., The reduction of *h'–f* records to electron-density-height profiles, *Penna. State Univ. Sci. Rept.*, **94,** 1957.

Shinn, D. H., and Whale, H. A., Group velocities and group heights from the magnetoionic theory, *J. Atmospheric Terrest. Phys.*, **2,** 85, 1952.

Thomas, J. O., Long, A. R., and Westover, D., The calculation of electron density profiles from topside sounder records, *J. Geophys. Res.*, **68,** 3237, 1963.

Titheridge, J. E., A new method for the analysis of ionospheric $h'(f)$ records, *J. Atmospheric Terrest. Phys.*, **21,** 1, 1961.

Unz, H. A., A solution of the integral equation $h'(f) = \int \mu'(f, f_0)\, dz(f_0)$, *J. Atmospheric Terrest. Phys.*, **21,** 40, 1961.

Warren, E. S., Sweep-frequency radio soundings of the topside of the ionosphere, *Can. J. Phys.*, **40,** 1692, 1962.

Wharton, C., *Proc. 4th Intern. Conf. Ionization Phenomena on Gases*, North-Holland Publishing Co., Amsterdam, 1959.

Calvert, W., and Goe, G. B., Plasma resonances in the upper atmosphere, submitted to *J. Geophys. Res.*, 1964.

ON THE HEIGHT OF AURORAL ABSORPTION

BENGT HULTQVIST*

Goddard Space Flight Center

The height distribution of the absorption produced by two different observed electron spectra, $N(E)\ dE=5\times10^{9}\ \mathrm{xp}(-E/5)\ dE$ and $N(E)\ dE=7\times10^{4}\ \exp(-E/41)\ dE$ and also by the bremsstrahlung from the first mentioned one has been estimated. The results indicate that the absorption in the 60–90-km range, due to hard electron spectra, often may dominate, whereas the absorption caused by bremsstrahlung is smaller than that due to the primary electrons producing the X-rays. These results are discussed in respect to the height of the absorbing layer and the observed very small variation of auroral absorption at sunrise and sunset.

INTRODUCTION

It was found in an early stage of the experimental investigation of the ionosphere by means of ionosondes that during aurora total blackout of the ionosonde often occurs, indicating strong absorption taking place below the reflecting layers. That the main part of the ionization producing the radiowave absorption that is characteristic for magnetically disturbed conditions in the auroral zone is located below the E-layer is evidenced also by several other observations. The few rockets containing electron density experiments that have been flown into the disturbed ionosphere in the auroral zone have all brought down information showing that absorption takes place well below the E-layer. Heikkila and Penstone (1961), for instance, found a pronounced peak in the height distribution of absorption per unit length at 75 km and no measurable absorption at all above 90 km. The rocket measurements of Seddon and Jackson, Kane and others (cf. e.g., Seddon and Jackson, 1958, and Kane, 1963) seem not to have given any results showing so high electron density up in the E-layer that the main part of the absorption might have taken place there. The Norwegian rocket measurements of electron density in the lowest ionosphere have shown strong increases in electron density in and below the normal D-layer height interval during aurora and magnetic storms (cf. Jespersen et al., 1963). Multifrequency riometer measurements often give equivalent heights below 75 km for the absorbing ionization (Ziauddin, 1963) and the time constants evaluated from absorption records indicate similar altitudes (Ziauddin, 1961a; Gustafsson, 1963). Furthermore, electron density profiles obtained from partial reflection and cross-modulation studies in the auroral zone show strong increases in the electron density, and the absorption caused by it, below 80 km during aurora and magnetic storms.

Some of the observational facts mentioned above led Chapman and Little (1957) to propose that bremsstrahlung X-rays, produced by the primary auroral electrons, are responsible for the major portion of the auroral absorption. However, the absorption caused by bremsstrahlung from primary auroral electrons, of the spectra observed by McIlwain (1960), is probably considerably less than that due to the primary electrons themselves (cf. Aikin and Maier, 1963, and below). Practically all the primary electrons found by McIlwain will be stopped above 80 km and most of the absorption takes place above 90 km.

Investigations of the variation of auroral absorption in the twilight periods (Hultqvist, 1962, 1963; Brown and Barcus, 1963; Holt and

*NASA—National Academy of Sciences—National Research Council Senior Postdoctoral Resident Research Associate on leave of absence from Kiruna Geophysical Observatory, Kiruna C, Sweden.

Landmark, 1963) have shown that the difference between postsunrise and presunset (day) absorption on one hand and presunrise and postsunset (night) values on the other hand is much less than expected on the basis of the present knowledge about the negative ions in the lowest ionosphere. The two most probable alternative interpretations of this observation have been proposed to be the following: either the main part of the ionization responsible for the auroral absorption is situated above 90 km altitude or the ratio of negative ion to free electron densities, λ, is much less than hitherto believed (Hultqvist, 1963b).

Campbell and Leinbach (1961) have calculated the absorption taking place in the height interval of the visual aurora from measured fluxes of auroral light and the ratio of ionization and excitation cross sections. They proposed that sometimes all auroral absorption may take place in the height interval where the visual form is located. On the basis of this and the absence of day-night variation in auroral absorption, Brown and Barcus (1963) concluded that auroral absorption usually takes place above 90 km altitude. This first one of the two above mentioned alternatives means, however, that all the observational evidence mentioned earlier for a low altitude of the absorbing ionization should be disregarded. This seems not very probable to this author.

While the rocket-measured auroral electron spectrum of McIlwain (1960) was very soft and produced the ionization mainly up in the altitude range of visible aurora (see below), some rocket experiments in the auroral zone have shown considerably harder electron-spectra. Davis et al. (1960) found the differential energy spectrum to be proportional to E^{-2}, while McDiarmid et al. (1961) observed an exponential spectrum with mean energy of 22 kev. In addition to these few isolated rocket observations of harder electron-spectra, there is now available results of satellite observations of similar high-energy electrons, obtained during many passages through the auroral zones (Mann et al., 1963). These observations support the low altitude absorption alternative. They will be discussed in this note and will first be briefly described.

OBSERVATIONS OF FLAT ELECTRON SPECTRA IN THE AURORAL ZONE

Mann, Bloom, and West (1963) flew magnetic spectrographs, covering the energy range 90–1200 kev for electrons in the Discoverer satellites Nos. 29 and 31. These satellites were launched on August 30 and September 17, 1961, respectively, in near polar orbits. Discoverer 29 had a perigee of 160 km near the North Pole and an apogee of 610 km. The perigee of 240 km was at 30°N for Discoverer 31 and its apogee was at 420 km. The opening angle of the instrument was small (about 2×4 degrees) and it was continuously directed outward along the radius vector from the earth's center.

Three different types of differential electron spectra were found. One proportional to $\exp(-E/5$ kev), i.e., the same as found by McIlwain (1960), was observed over the polar caps and sometimes down to rather low latitudes, especially in the northern hemisphere. The fluxes of these steep spectrum electrons were sometimes 10–100 times greater than that found by McIlwain (1960) to be associated with a weak aurora of international brightness coefficient (IBC) about I. The differential energy spectrum obtained by McIlwain was:

$$N(E)=5\times10^{8}\exp(-E/5\text{ kev})$$

$$\text{electrons cm}^{-2}\text{sec}^{-1}\text{ster}^{-1}\text{kev.}^{-1} \quad (1)$$

In the calculations below, a 10 times more intense flux will be employed.

A second type of spectrum was observed only in the region of the South Atlantic magnetic anomaly and could be interpreted as the lowest tail of the VanAllen belt.

The third type of spectrum observed is the one that is of most interest here. It was a class of harder electron spectra than that reported by McIlwain (1960), predominantly found in the auroral zones. These electrons were observed when the spectrometer looked within 10–20 degrees of the geomagnetic field lines. This indicates that the electrons were injected into the atmosphere. They were probably primary auroral electrons, according to Mann et. al. (1963).

The spectra of this kind observed during 20 different passages through the auroral zones could

be grouped into two classes with regard to energy fall-off: one given by $N_0 \exp(-E/25 \text{ kev})$ and the other by $N_0 \exp(-E/42 \text{ kev})$. Some of the spectra, however, had e^{-1} energy values as low as 15 kev and in one case a high value of 165 kev was found.

The calculations below will be made for a differential electron energy spectrum of

$$N(E) = 7 \times 10^4 \exp(-E/41)$$
$$\text{electrons cm}^{-2}\text{sec}^{-1}\text{ster}^{-1}\text{kev}^{-1} \quad (2)$$

obtained from figure 5 in the report of Mann et. al. (1963).

There was a magnetic storm when Discoverer 29 was in orbit and it was mainly in the course of that storm that the hard electron spectra were observed in the auroral zones. There was probably a time lag between the start of the storm and the onset of the electron bombardment, but due to incomplete orbit coverage nothing definite could be stated. The electron flux definitely diminished as the storm waned.

ABSORPTION PRODUCED BY PRIMARY AURORAL ELECTRONS, GENERAL

The exact calculation of the energy dissipation of fast electrons in the atmosphere is very complicated because of the range straggling due to large single-energy losses occurring in both radiative and inelastic collisions and also because of the large angular deviations encountered by the electrons in elastic collisions. Even the extensive numerical calculations of Spencer (1959) give somewhat inaccurate results for large thicknesses of the absorbing material.

Maeda (1963) has given the following empirical relation for the attenuation of electrons with energies between 5 and 300 kev:

$$N(E,x)\, dE = N_0(E) \exp(-0.318 \times 10^7 E^{-2.2} x)\, dE \quad (3)$$

where $N_0(E)$ is the initial differential intensity of electrons with a kinetic energy of E, and x is given in g/cm^2. This expression is based on laboratory measurements and takes into account elastic scattering. As soon as an electron has undergone interaction, so its energy is outside the interval dE at E, it is considered as lost from the beam. In using expression 3 for calculation of the energy dissipation, one thus considers the total energy of an electron that has been inelastically scattered at an atmospheric depth, x, as dissipated at that same depth. This gives some overestimation of the energy dissipation at small atmospheric depths and thus produces the ionized layer at somewhat too high an altitude in the atmosphere. On the other hand, since the electron flux is certainly not attenuated in an exponential way close to the electron range, it seems likely that an overestimation is made also of the very lowest part of the produced ionization, when formula 3 is employed. It is, however, probable that the errors in the height distribution of the produced ionization are not large, measured in km, in the atmosphere where the density increases approximately exponentially with decreasing height. More on this in the discussion on page 704.

Expression 3 has the great advantage of making all calculations easy. It will be used below for the estimates of the absorption caused by various electron spectra. The geomagnetic field lines will be assumed to be vertical in the auroral zones.

The flux of electrons with a pitch angle α at the atmospheric depth x g/cm^2 is then given by

$$2\pi N(\alpha,x,E)\, dE \sin\alpha\, d\alpha = 2\pi N_0(E,\alpha) \exp[-x/\sigma(E)\cos\alpha] \sin\alpha\, d\alpha\, dE \quad (4)$$

where $1/\sigma\ (E) = 0.318 \times 10^7\ E^{-2.2}$ and $N_0(E,\alpha)$ is the differential electron spectrum outside the atmosphere. Based on the experimental results reviewed above, it will be assumed that the low energy spectrum $[\alpha \exp(-E/5)]$ is isotropic over the upper hemisphere, but that the high energy electrons $[\alpha \exp(-E/41)]$ come in only within one steradian around the field lines, and they will be assumed to propagate vertically.

(a) Isotropic Flux

Integrating the pitch angle from 0 to $\pi/2$ and substituting $y = x/\cos\alpha$ we obtain

$$N(x,E)dE = 2\pi N_0(E) G[x/\sigma(E)]\, dE \quad (5)$$

where $G[x/\sigma(E)]$ is the so-called Gold integral (cf., e.g., Rossi, 1952):

$$G[x/\sigma(E)] = \int_1^\infty \exp[-xs/\sigma(E) s^{-2}\, ds. \quad (6)$$

$$dN(x,E)/dx = 2\pi N_0(E) \cdot dG/dx = 2\pi [N_0(E)/\sigma(E)] \text{Ei}[-x/\sigma(E)] \quad (7)$$

where $\mathrm{Ei}[-x/\sigma(E)]$ is the exponential integral, defined by

$$-\mathrm{Ei}(-y)=\int_y^\infty e^{-z}z^{-1}\,dz. \tag{8}$$

The energy dissipation rate to the atmosphere per unit volume by electrons of energy E is given by

$$-\rho(h)E\cdot dN(E,x)/dx$$
$$=2\pi\rho(h)E[N_0(E)/\sigma(E)]\mathrm{Ei}[-x/\sigma(E)]$$
$$\text{kev cm}^{-3}\text{ sec}^{-1}\text{ kev}^{-1} \tag{9}$$

where E is measured in kev and h is the altitude in cm. By taking the average amount of energy used in production of one electron-ion pair equal to 32 ev, $N_0(E)=5\times10^9\exp(-E/5)$ electrons cm^{-2} $\text{sec}^{-1}\text{ster}^{-1}\text{kev}^{-1}$ and $1/\sigma(E)=0.318\times10^7E^{-2.2}$, we obtain the electron production rate $q(E,h)$:

$$q(E,h)=-3.12\times10^{18}\cdot\rho(h)E^{-1.2}$$
$$\times e^{-E/5}\,\mathrm{Ei}(-0.318\times10^7E^{-2.2}x)$$
$$(\text{cm}^3\text{ sec kev})^{-1}. \tag{10}$$

$-\mathrm{Ei}(-x)$ goes to infinity when x goes to zero. The expression 10 is valid when the isotropic flux outside the atmosphere has infinite extension in the horizontal plane. In practice this means that the auroral electron bombardment must be homogenous over areas several hundred kilometers in extension for (10) to give correct results for small atmospheric thicknesses. This condition is certainly not fulfilled in most auroras. However, it only influences the dissipation in the highest layers of interest, making the electron density larger. The absorption is in any case small there. In the height interval where most of the energy dissipation takes place expression 10 should be accurate from that point of view.

(b) Vertical Incidence

For vertical incidence the expression for the energy dissipation rate to the atmosphere per unit volume by electrons of energy E is given by

$$-E\cdot dN(E\cdot h)/dh=\rho(h)\cdot E\cdot[N_0(E)/\sigma(E)]$$
$$\times\exp[-x/\sigma(E)]\text{Kev}(\text{cm}^3\text{ sec Kev})^{-1} \tag{11}$$

and the electron production rate, $q(E,\ h)$, for $N_0(E)=7\times10^4\exp(-E/41)$ electrons $\text{cm}^{-2}\text{ sec}^{-1}$ kev^{-1}, by

$$q(E,h)=0.695\times10^{13}\,\rho(h)E^{-1.2}$$
$$\exp(-E/41)\exp(-0.318\times10^7E^{-2.2}x)$$
$$\text{electrons (cm}^3\text{ sec Kev)}^{-1}. \tag{12}$$

ABSORPTION DUE TO PRIMARY AURORAL ELECTRONS AND BREMSSTRAHLUNG

With the use of equation 10 the electron production rate $q(E,\ h)$, due to a steep spectrum of McIlwain's type but with a 10 times higher flux (as found by Mann et al., 1963; it corresponds to an aurora of IBC II) was computed for every 10th Kev from 5 Kev up to 65 Kev, and for every 10th km between 70 and 130 km. The total electron production rate due to the complete spectrum was obtained by numerical integration.

The stationary state electron density, N_e, was derived from

$$N_e(h)=[q/(1+\lambda)\cdot(\alpha_d+\lambda\alpha_n)]^{1/2} \tag{13}$$

The profiles used by Nicolet and Aikin (1960), Aikin (1962), and others, and the values $\alpha_d=4.6\times10^{-7}\text{ cm}^3\text{ sec}^{-1}$ and $\alpha_n=10^{-7}\text{ cm}^3\text{ sec}^{-1}$ for the dissociative (α_d) and ion-neutralization (α_n) recombination coefficients have been used. Finally the absorption per km height interval was computed for the riometer frequency 27.6 Mc/s in the auroral zone, using the old Appleton-Hartee

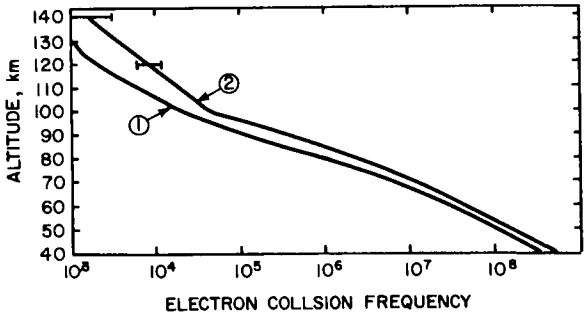
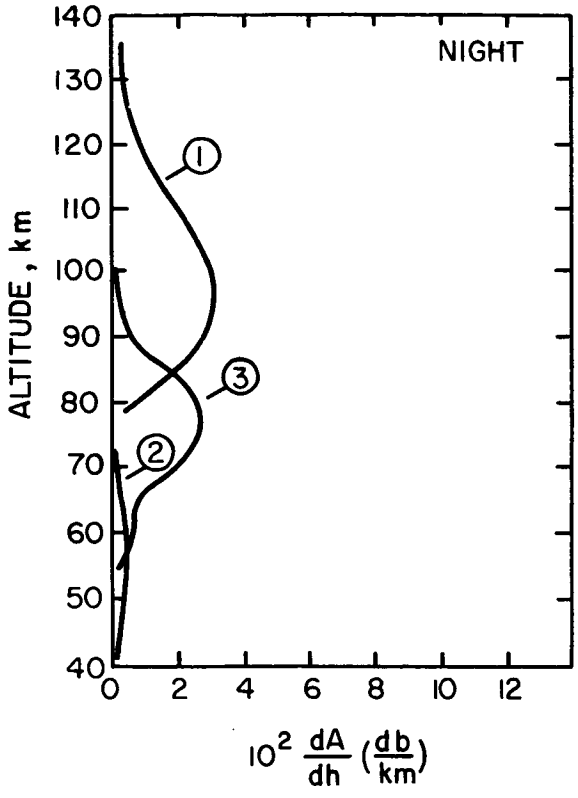
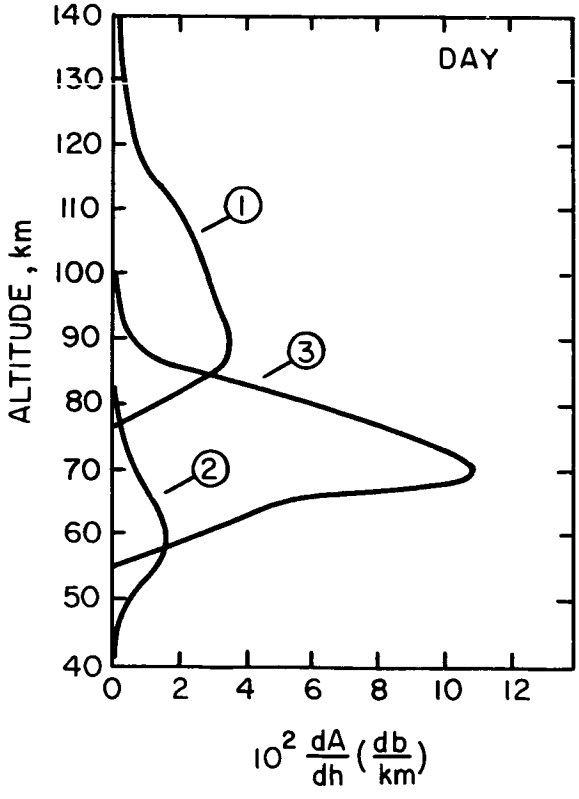


FIGURE 1.—The electron collision frequency, ν, as a function of altitude. Curve 1 has been drawn on the basis of data presented by Holt (1963) for the lower half of the altitude range. In the upper half curve 1 has been extrapolated so that it parallels the curve shown by Ratcliffe and Weekes (1960) in about the same way as in the lower altitudes. Curve 2 is after Nicolet (1959) below 90 km and after Hanson (1961) above 100 km.





expression

$$dA/dh = 0.459 \times 10^5 N_e \nu / (3.34 \times 10^{16} + \nu^2) \text{ (db/km)}. \quad (14)$$

Computations were made for the two electron collision frequency profiles shown in figure 1. ν-profile No. 1 is probably more represeatetive for the actual ionospheric situation at the lower heights, where it is based on recent measurements (Holt, 1963). With the use of ν-profile no. 2 in figure 2, absorption-per-km values higher by a factor of about two is obtained and the ratio between the high-altitude and low-altitude absorption contributions is also a little changed, but the differences are not of major importance here. Below, only the results obtained with the use of ν-profile No. 1 will be presented. The probable inaccuracy of the numerical absorption values will be discussed in some more detail later.

The result of the computations is shown as curves nos. 1 in figure 2a (for daytime) and 2b (for night). The total absorption amounts to 1.0 db in the day and to 0.9 at night. dA/dh has its maximum at about 90 km during the day and at 95 km by night. Most of the absorption takes place above 90 km.

Aikin and Maier (1963) have calculated the electron production rate due to the bremsstrahlung resulting from the electron spectrum measured by McIlwain (1960) during an aurora of IBC I. Their electron production rates, multiplied by a factor of 10 to make them correspond to the absorption given by curves Nos. 1 in figure 2 have been converted into absorption per km in the way described above. The result is shown as curves Nos. 2 in figures 2, a and b.

FIGURE 2.—(a) is for daytime and (b) for night. Curves 1 in (a) and (b) show the height distribution of the absorption produced by the differential energy spectrum $N(E) = 5.10^9 \exp(-E/5 \text{ Kev})$ electrons cm^{-2} sec^{-1} ster^{-1} Kev^{-1}. The total absorption values corresponding to curves 1 amount to 1.04 db in the day and to 0.89 db at night. Curves 2 give the absorption due to the bremsstrahlung of the same electron spectrum. Total absorption in the day is 0.27 db and in the night 0.061 db. Curves 3, finally, represent the absorption distribution produced by the differential energy spectrum $N(E) = 7.10^4 \exp(-\text{E}/41 \text{ Kev})$ electrons cm^{-2} sec^{-1} Kev^{-1}, coming in along the field lines. Total daytime absorption is 1.9 db and the nighttime one is 0.52 db.

The total daytime absorption due to bremsstrahlung amounts to 0.27 db, or about one-fourth of the absorption produced by primary electrons. It has its maximum at about 60 km. The thickness of the layer at its half-value points is 17 km. At night the total bremsstrahlung absorption is very small, 0.06 db, i.e., about one-fifteenth of the corresponding absorption due to primary auroral electrons.

Finally the absorption due to the hard electron spectrum $N(E)\ dE = 7\times10^4 \exp(-E/41 \text{ kev})\ dE$, observed by Mann et. al. (1963), has been computed using expression 12 for the electron production rate and the same procedure as described earlier for transferring $q(E,h)$ into dA/dh. The result is shown in figure 2, a and b, as curves Nos. 3. The integration was carried out between 50 and 300 Kev. For the spectrum above, electrons with energies less than 50 Kev give negligible contribution to the absorption.

The total daytime absorption was found to be 1.9 db and the nighttime one 0.5 db. The height of maximum dA/dh was 70 km in the day and a little less than 80 km at night. Thus the absorption in the 60- to 90-km height interval, due to the hard electrons, is twice as large as that due to the soft ones and one order of magnitude greater than that produced by bremsstrahlung in the day.

DISCUSSION

The uncertainty in the numerical values presented in figure 2 is large. The inaccuracy in the knowledge of the collision frequency introduces a possible error in the absorption of a factor of two or even more at the highest levels shown in the figure. The density in the upper half of the altitude range in figure 2 may vary appreciably with local time and season. This together with the low degree of accuracy in the existing experimental density values for this region, makes an uncertainty of a factor of two possible. In addition, the use of the classical magnetoionic formula result in values of dA/dh which, in certain atmospheric depths, may be 50% too low (cf., e.g., Hultqvist, 1963c).

It has been mentioned above that the calculation of the electron production rate due to an isotropic electron flux presumes a very large extension of the area of electron influx into the atmosphere. The area should be so large that electrons entering at almost horizontal direction at one side are stopped after having travelled a small part of the distance to the other side. That means that the dimensions of the area of electron influx has to be many hundred kilometers. It is certainly not true in nature that the electron flux is homogeneous over such areas. It is known, for instance, that the correlation of auroral absorption records decreases to 0.5 for a distance between the riometers of 300–400 km (Holt et al., 1961). The result of the assumption is that the absorption in the upper levels is overestimated. However, the influence on the total radio wave absorption is probably quite small, as the absorption at fairly great atmospheric depths dominates strongly. The effect on the altitude of the absorbing layer of using the simple expression 3 is certainly more important.

It has been shown by Young (1956) that the average energy lost per electron by a beam of electrons of energies from a few Kev to a few tens of Kev in traversing an absorber is approximately equal to the energy, E_0, of an electron with an end-point range equal to the absorber thickness. This was used by McIlwain (1960) in deriving the spectrum of the electrons observed by him in aurora. McIlwain pointed out that if the integral number energy spectrum can be represented by a function of the form $c \exp(-E/b)$, where c and b are constants, then the energy flux emerging from an absorber with an electron end-point range energy of E_0 will be $cb \exp(-E_0/b)$. Using this and the empirical relations between energy and practical range, R, for monoenergetic low-energy electrons by Katz and Penfold (1952), the following expression can be derived for the electron production rate, $q(h)$ (el./cm^3 sec), due to an electron flux which is isotropic outside the atmosphere

$$q(h) = \frac{2\pi c}{0.032}\,\rho \int_0^{\pi/2} \frac{E_0(\theta)\exp[E_0(\theta)/b]\sin\theta\, d\theta}{R(\theta)\,[2.853-0.191 \ln E_0(\theta)]} \tag{15}$$

where $R(\theta)$ can be taken equal to $1.36p$ (mm Hg/$\cos\theta$ for the height interval of interest here. Numerical calculations have shown that the equilibrium electron density due to the soft spectrum $n(E) = 5\times10^9 \exp(-E/5)$ electrons/cm^2

sec ster Kev obtained with the use of (15) differs by less than 50% from those found by means of Maeda's (1963) formula 3, except at the bottom of the ionized layer (90 km) where the difference amounts to almost 100%. Those results indicates that the use of Maeda's attenuation formula is not probable to involve errors in the absorption per unit height interval larger than a factor of two in any part of the height interval of interest, except possibly in the very low tail. The effect of this on ratios between total absorption due to various sources or on day to night ratios is probably small compared with other uncertainties.

An assumption of infinite extension in the horizontal plane of the area of electron influx was also made by Aikin and Maier (1963) in their calculation of the ionization produced by bremsstrahlung. Since the bremsstrahlung photons can travel very far in almost horizontal direction in the upper levels of interest here, the required extension of the area of influx is still higher than for the electrons. The effect of this assumption is probably that the calculated ionization rate, due to bremsstrahlung, is too high. It is difficult to give quantitative values of this overestimation. In addition to this uncertainty in the calculation of the bremsstrahlung ionization rate, there are errors introduced by the specific approximations and simplifications made by Aikin and Maier in deriving the expressions for the X-ray flux and its absorption in the atmosphere.

The conclusion of this discussion of the accuracy of the absorption values given in figure 2 is that only the order of magnitude is significant. The ratios of the absorption values for the various ionization sources and for the highest and lowest altitudes in figure 2 are probably correct within a factor less than four.

With the uncertainties in some of the parameters, of the order of magnitude mentioned, extensive calculations giving high degree of accuracy in other parameters, seem not to be justified. It may be of interest to mention that an absorption-height distribution calculated on the assumption that the electron flux consists only of the electrons within one steradian around the field-line direction, propagating strictly along the field lines, instead of being isotropic over 2π steradians, is only about 50% smaller than that obtained with isotropic flux. The heights of the absorbing layers and the shapes are similar in both cases. It is therefore quite reasonable to make the simplifying assumption of vertical influx for a rough estimation.

(2) It can be seen in figure 2 that the absorption deep in the atmosphere, due to bremsstrahlung, is probably much less than that produced mainly above 85 km by the primary electrons, for the steep electron spectra found by McIlwain. For decreasing steepness of the electron spectrum one would expect an increasing importance of the absorption produced by bremsstrahlung because of the increasing cross section for bremsstrahlung production with electron energy. On the other hand, when the primary electrons become more energetic, they ionize lower down in the atmosphere; the height difference between the ionized layers produced by primary electrons and the bremsstrahlung decreases and with it the difference in absorption cross section for the electrons in the two layers. This tends to make the bremsstrahlung less important. More calculations for various spectra are needed before a general statement about the importance of the bremsstrahlung in producing radio-wave absorption can be given, but it seems probable that the high altitude absorption due to primary electrons—and to heavy auroral particles—is greater than the absorption produced by bremsstrahlung for all spectra of interest.

(3) For the parameter values used in the calculations leading to the curves 1 and 2 in figures 2a and b, we expect 1 db absorption at 27.6 Mc/s for an aurora of IBC II. To obtain the corresponding absorption values for the same electron energy spectrum for IBC I we only have to divide by the square root of 10, as there is a factor of 10 for the light emission, and therefore for ionization rate, between each IBC value. Thus an IBC I aurora (in the night) would only give 0.3 db due to primary electrons and 0.02 db due to bremsstrahlung for a McIlwain spectrum. For an IBC III aurora the corresponding values would be 3 db and 0.2 db, respectively. An IBC IV aurora, finally, would give 10 db due to primary electrons and 0.6 db because of bremsstrahlung.

(4) Figure 2 shows that the absorption produced at low altitudes by the flat electron spectra

observed by Mann et al. (1963), may be more important than that due to the very steep spectra, also observed by Mann et al. as well as by McIlwain. The data published by Mann et al. (1963), do not give too much information about the range over which the hard electron flux varies. Comparison of curve 3 in figure 2 and the absorption to be expected for auroras of various IBC, produced by electrons with spectra of McIlwain's type, show that the daytime absorption under curve 3 dominates over that due to low-energy electrons for IBC I and II (fig. 2 illustrates the situation for an IBC 2 aurora) but not for III and IV. For nighttime (fig. 2b) the absorption due to high-energy electrons is greater than that produced by the low-energy ones only for IBC I aurora (with the hard electron flux unchanged), while for IBC II it is somewhat smaller. It seems, however, reasonable to assume that also the flux of electrons with flat spectrum varies at least a factor of 10 up and down from that used in this note, and that its daily average value is correlated with the daily flux of steep spectrum electrons. That this is so is supported by the results of balloon observations of X-rays. The absorption due to these high-energy electrons will then be the dominating one in the average.

Even if there should exist a statistical relation between the fluxes of steep-spectrum and flat-spectrum electrons, it is most probable that wide variations in the resulting total spectrum can be found from one aurora to another. One can expect to see both auroras (weak ones) without appreciable absorption and strong absorption without visible aurora. Such combinations can be obtained from the two spectral types dealt with in this note. The dissimilarity between the diurnal variation curves for visual aurora and auroral absorption may be understood on this basis.

From the report of Mann et al. (1963), one does not get information about the occurrence frequency of the two above-mentioned types of spectra. The data from Injun I show that the radio-wave absorption is well correlated with the flux of electrons of energy greater than 40 Kev (Machlum and O'Brien, 1963). This, as well as the good correlation between radio-wave absorption and fluxes of bremsstrahlung X-rays of energies up towards 100 Kev, indicates that auroral type of absorption is not primarily caused by electrons with the very steep spectrum found by McIlwain (1960). The rocket observation of Heikkila and Penstone (1961), on the other hand, can be understood as an effect of a flat-electron spectrum without any contribution from the steep type.

(5) The day-to-night ratio of the absorption, A_D/A_N, obtained from figure 2a and b, is, for absorption due to low-energy electrons alone, 1.12; for the absorption produced by bremsstrahlung, 4.4; for absorption due to low-energy electrons *and* bremsstrahlung, 1.34; for high-energy electron absorption, 3.7; and finally, for the sum of absorption caused by low and high energy electrons and by bremsstrahlung, 2.2.

The observed A_D/A_N value for auroral absorption is between one and two (1.1–1.2 according to Hultqvist, 1962, 1963; about 1 according to Brown and Barcus, 1963, about 2 according to Holt and Landmark, 1963). Probably it is fairly close to unity.

The absorption values shown in figure 2 were calculated with the use of the height distribution of λ (the ratio between negative ion and free electron densities) used by Nicolet, Aikin and others in the last few years.

Of the A_D/A_N values obtained from figure 2, which were given above, only the "low-energy electron" one agrees with the observed value of A_D/A_N. The experimental values of A_D/A_N were obtained by averaging over all auroral absorption events recorded over extended periods of time ($3\frac{1}{2}$ years in Hultqvist's case).

It was mentioned earlier that there is experimental evidence showing that auroral absorption is usually not caused by low-energy electrons alone. When assuming equal probability of occurrence of the steep- and flat-electron energy spectra, which seems to be a reasonable assumption in absence of detailed statistical data, one would expect an A_D/A_N value of more than two, as mentioned earlier.

It should be mentioned here that the large uncertainty in the absolute absorption values does not affect the A_D/A_N ratio too much. The A_D/A_N value is more dependent on the height distribution. The larger the fraction of the total absorption that takes place above 90 km, the

smaller the A_D/A_N value will be. From the discussion earlier, it can be concluded that it is more probable that the calculated high-altitude absorption is overestimated than that it is too small.

On the basis of what has been said above, it seems possible to conclude that there is a significant discrepancy between the A_D/A_N values calculated on the basis of the λ profiles of Nicolet and others, on the one hand, and the observed values on the other. This means that the observed absence of day-night variation in auroral absorption is *not* due to the main part of the absorbing ionization being located above 90 km. The second alternative mentioned in the introduction—that of the λ profiles being lower than believed before—is therefore supported by the results discussed in this note.

ACKNOWLEDGMENT

I am grateful to L. R. Davis and D. S. Evans for valuable discussions.

REFERENCES

Aikin, A. C., *J. Atm. Terr. Phys.*, **23,** 287–300 (1962).

———, and Maier, E. J., The effect of auroral bremsstrahlung on the lower ionosphere. Paper presented at the NATO Advanced Study Institute, Skeikampen, Norway, April 17–26, 1963.

Brown, R. R., and Barcus, J. R., Day-night ratio for auroral absorption events associated with negative magnetic bays, *J. Geoph. Res.*, **68,** 4175–80, (1963).

Campbell, W. H., and Leinbach, H., Ionospheric absorption at times of auroral and magnetic pulsations, *J. Geoph. Res.*, **66,** 25–34 (1961).

Chapman, S., and Little, C. G., The nondeviative absorption of high frequency radio waves in auroral latitudes, *J. Atm. Terr. Phys.*, **10,** 20 (1957).

Davis, L. R., Berg, O. E., and Meredith, L. H., Direct measurements of particle fluxes in and near auroras, *Space Research*, pp. 721–735, North-Holland Publishing Co., Amsterdam (1960).

Gustafsson, G., Ionization in the D-region during auroral breakup events. Submitted to Planet. Space Sci. for publication (1963).

Hanson, W. B., Structure of the ionosphere, Satellite environment handbook, pp. 27–46, Stanford University Press, Stanford, 1961.

Heikkila, W. J., and Penstone, S. R., Rocket measurements of auroral radio absorption, *Can. J. Phys.*, **39,** 1875–81 (1961).

Holt, O., Landmark, B., and Lied, F., NDRE Report No. 35, parts I to III, Norwegian Defence Research Establishment, Kjeller, April 1961.

———, Some experimental studies of the ionospheric D-region at high latitudes, NDRE Report No. 46, Norwegian Defence Research Establishment, Kjeller, July 1963.

———, and Landmark, B., Radio astronomical and satellite studies of the atmosphere, pp. 251–222, North-Holland Publishing Co., Amsterdam, 1963.

Hultqvist, B., Kiruna Geophysical Observatory, Scientific Report No. 2, contract No. AF 61(052)-601, May 1962.

———, Sunrise and sunset effects on cosmic noise absorption associated with aurora and magnetic storms, *Planet Space Sci.*, **11,** 371–383 (1963).

———, On the height distribution of the ratio of negative ion and electron densities in the lowest ionosphere, *J. Atm. Terr. Phys.*, **25,** 225–240 (1963b).

———, Studies of ionospheric absorption of radio waves by the cosmic noise method. Radio astronomical and satellite studies of the atmosphere, pp. 163–219, North-Holland Publishing Co., Amsterdam, 1963c.

Jespersen, M., Petersen, O., Rybner, J., Bjelland, B., Holt, O., and Landmark, B., Electron density and collision frequency observations in the lower D-region during auroral absorption, Report No. 16, Ionosphere Laboratory, Royal Technical University of Denmark, Copenhagen, March 1963.

Kane, J. A., Electron densities in the E-region deduced from rocket observations, Electron Density Profiles, pp. 67–69, The MacMillan Co., New York, 1962.

Katz, L., and Penfold, A. S., Range-energy relations for electrons and the determination of beta-ray end-point energies by absorption, *Rev. Mod. Phys.*, **24,** 28–44 (1952).

Maeda, K., Auroral dissociation of molecular oxygen in the polar mesosphere, *J. Geoph. Res.*, **68,** 185–197 (1963).

Maehlum, B., and O'Brien, B. J., Study of energetic electrons and their relationship to auroral absorption of radio waves, *J. Geoph. Res.*, **68,** 997–1010 (1963).

Mann, L. G., Bloom, S. D., and West, H. I., Jr., The electron spectrum from 90 Kev to 1200 Kev as observed on Discoverer satellites 29 and 31, *Space Research III*, pp. 447–462, North-Holland Publishing Co., 1963.

McDiarmid, I. B., Rose, D. C., and Budzinski, E., Direct measurement of charged particles associated with auroral zone radio absorption, *Space Research II*, pp. 1194–1205, North-Holland Publishing Co., Amsterdam (1961).

McIlwain, C. E., Direct measurement of particles producing visible auroras, *J. Geoph. Res.*, **65,** 2727–2747 (1960).

Nicolet, M., Collision frequency of electrons in the terrestrial atmosphere, *Phys. Fluids*, **2,** 95–99 (1959).

———, and Aikin, A. C., The formation of the D-region of the ionosphere, *J. Geoph. Res.*, **65,** 1469–1483 (1960).

RATCLIFFE, J. A., and WEEKES, K., The ionosphere, *Physics of the upper atmosphere*, pp. 377–470, Academic Press, New York (1960).

ROSSI, B., High energy particles, Prentice-Hall, New York (1952).

SEDDON, J. C., and JACKSON, J. E., *Ann. Geophys.*, **14,** 456 (1958).

SPENCER, L. V., Energy dissipation by fast electrons, National Bureau of Standards Monograph 1, Sept. 10, 1959.

WINCKLER, J. R., BHAVSAR, P. D., and ANDERSON, K. A., A study of the precipitation of energetic electrons from the geomagnetic field during magnetic storms, *J. Geoph. Res.*, **67,** 3717–3736 (1962).

YOUNG, J. R., Penetration of electrons and ions in aluminum, *J. Appl. Phys.*, **27,** 1–4 (1956).

ZIAUDDIN, S., Thesis at the University of Saskatchewan (1961).

———, *Nature*, **191,** 1984 (1961a).

EVIDENCE FOR THE EFFECT OF CORPUSCULAR RADIATION ON THE IONOSPHERE

F. MARIANI*

Goddard Space Flight Center

The correlation of maximum electron density in the F region at noon with solar activity is investigated for both long-term and month-to-month variations, in data for 1937–57 from all existing observatories. On the basis of noon-equilibrium between electron loss and ion production by solar radiation, an ionizing effect, strongly dependent on solar activity and peaked at latitudes of 55° to 65° is shown to exist.

This latitudinal variation suggests a corpuscular origin of the effect. The possible source of a corpuscular flux is discussed. The energy of the ionizing particles is in the range of Kev. Quantitative evidence is given that Van Allen belts could be an important, or possibly the main, source.

1. INTRODUCTION

Recent studies (Mariani, 1959 and 1960, here indicated as M1 and M2) on the variations of electron density distribution in the F2 layer suggest some latitudinal and 11-year variations of the solar ultraviolet radiation. The interpretation of these results is rather difficult because the experimental data can be affected by other than solar causes (local variations of atmospheric temperature and winds, diffusion, etc.). Moreover, in the study of long series of ionospheric data, one cannot completely eliminate the "regular" effect due to the seasonal variation of zenith distance of the sun. Because of these difficulties, it is desirable to apply statistical methods to the data of all existing observatories. Local irregularities and systematic effects in a number of observatories are then easily recognizable, and some general conclusions can be drawn.

It is the purpose of this paper to apply a statistical analysis to all ionospheric data collected in the two 11-year periods, 1937–1947 and 1947–1957.

2. EXPERIMENTAL DATA AND METHOD OF ANALYSIS

In this study, we used the monthly median values of f_0 F2 at noon for about 70 observatories, listed in table 1, whose data are published by the Central Radio Propagation Laboratory of the National Bureau of Standards or directly by the observatories.

For indicators of solar activity, we used the monthly means of final sunspot numbers and of sunspot areas, published by the observatories of Zurich and Washington, respectively; the monthly mean areas of hydrogen filaments and hydrogen and calcium flocculi, deduced from the character numbers published by the Astro-physical Observatory of Arcetri; and the monthly values of heliographic distribution of chromospheric activity published by the observatory of Meudon. Details on the above parameters of solar activity are given elsewhere. The method of analysis is the

*NAS–NASA senior postdoctoral research associate, on leave of absence from the University of Rome.

TABLE 1.—*Locations of Observatories Used, in Geomagnetic and Geographical Coordinates*

Observatory	Geomag. coordinates		Geograph. coordinates	
	Lat.	Long. (east)	Lat.	Long. (east)
Clyde	82N	1	70N	291
Resolute Bay	82N	289	75N	265
Godhavn	80N	32	69N	306
Baker Lake	74N	315	64N	264
Narsassuaq	71N	38	61N	315
Reykjavik	70N	71	64N	338
Churchill	69N	323	59N	266
Point Barrow	68N	241	71N	203
Tromso	67N	117	70N	19
Kiruna	65N	116	68N	20
Fairbanks (col.)	65N	256	65N	148
Anchorage	61N	258	61N	210
Inverness	61N	83	57N	356
Oslo	60N	11	60N	10
Winnipeg	60N	323	50N	263
Uppsala	59N	106	60N	18
St. John	59N	21	48N	307
Prince Rupert	58N	283	54N	230
Ottawa	57N	351	45N	284
Leningrad	56N	118	60N	31
Slough	54N	83	51N	359
De Bilt	54N	89	52N	5
Lindau	52N	94	52N	10
Moscow	50N	121	55N	37
Freiburg	50N	90	48N	8
Washington	50N	350	39N	283
Poitiers	49N	82	47N	0
Schwarzenburg	48N	89	47N	7
Adak	47N	240	52N	183
Graz	47N	97	47N	15
Tomsk	45N	160	56N	85
San Francisco	44N	298	37N	238
White Sands	41N	316	33N	253
Baton Rouge	41N	334	30N	269
Casablanca	38N	69	34N	352
Wakkanai	35N	206	45N	142
Alma Ata	33N	152	43N	77
Portorico	30N	2	18N	293
Akita	29N	205	40N	140
Tokyo	25N	206	36N	140
Yamagawa	21N	198	31N	131
Dakar	21N	55	14N	343
Maui	21N	268	21N	203
Panama	21N	348	9N	280
Delhi	19N	149	29N	77
Okinawa	15N	196	26N	128
Formosa	14N	189	25N	121
Ibadan	11N	75	7N	4
Bombay	10N	144	19N	73
Djibouti	7N	114	11N	43
Baguio	5N	189	16N	121
Madras	3N	150	13N	80

TABLE 1.—Continued

Observatory	Geomag. coordinates		Geograph. coordinates	
	Lat.	Long. (east)	Lat.	Long. (east)
Guam	3N	212	13N	145
Tiruchy	1N	148	11N	79
Huancayo	0.6S	354	12S	285
Leopoldville	3S	84	4S	15
Singapore	10S	173	1N	104
Rarotonga	21S	274	21S	200
Buenos Aires	23S	9	35S	302
Tananarive	24S	113	19S	48
Johannesburg	27S	91	26S	28
Capetown	33S	80	34S	18
Brisbane	36S	227	27S	153
Falkland Island	40S	9	52S	302
Watheroo	42S	186	30S	116
Canberra	44S	225	35S	149
Christchurch	48S	253	44S	173
Hobart	52S	225	43S	147
Decepcion	52S	7	63S	299
Port Lockroy	53S	4	65S	297
Campbell Island	57S	253	53S	169
Macquarie Island	61S	243	54S	159
Terre Adelie	75S	231	67S	140

same as that used in the previous papers M1 and M2.

Following is a list of symbols:

N = monthly median values of $(f_0\ \mathrm{F2})^2$, in $(\mathrm{Mc/s})^2$
N_{12} = seasonal (12 month) variation of N
R = monthly mean of sunspot number
A_R = monthly mean of sunspot area
A_F = monthly mean of hydrogen-filament area
$A_{\Phi\mathrm{H}}$ = monthly mean of hydrogen-flocculi area
$A_{\Phi\mathrm{Ca}}$ = monthly mean of calcium-flocculi area
A_i = one of the above quantities R, A_R, A_F, etc.

When the above symbols have a bar (for example $\bar{N}$, $\bar{A}_F$, etc.) they represent the calculated corresponding value for the long-term (11-yr) variation.

3. THE CORRELATION OF F2 LAYER MAXIMUM ELECTRON DENSITY AT NOON WITH SOLAR ACTIVITY

We considered for each observatory simple and double linear regressions of the maximum electron density at noon with one solar parameter A_i (or two different parameters, A_i and A_j) expressed in the analytical form

$$\bar{N} = \bar{N}_0(1+\bar{\alpha}_i\bar{A}_i) \tag{1}$$

$$\bar{N} = \bar{N}_0(1+\bar{\beta}_i\bar{A}_i+\bar{\gamma}_j\bar{A}_j) \tag{2}$$

for long-term variations or,

$$N-N_{12} = (N-N_{12})_0(1+\alpha_iA_i) \tag{3}$$

$$N-N_{12} = (N-N_{12})_0(1+\beta_iA_i+\gamma_jA_j) \tag{4}$$

for actual month-to-month variations.

The quantities $\bar{N}_0$ and $(N-N_{12})_0$ represent the values deduced from the secular variation and from the month-to-month variation $N-N_{12}$ at minimum solar activity ($A_i=0$), i.e., the electron density for an absolutely quiet sun.

In M1 we found for the years 1947–1954 a remarkable dependence of $\bar{N}$ and $N-N_{12}$ upon the areas A_F of hydrogen filaments, after eliminating the dependence upon the sunspot number; a rough confirmation of this result was obtained in M2 from the few available data corresponding to

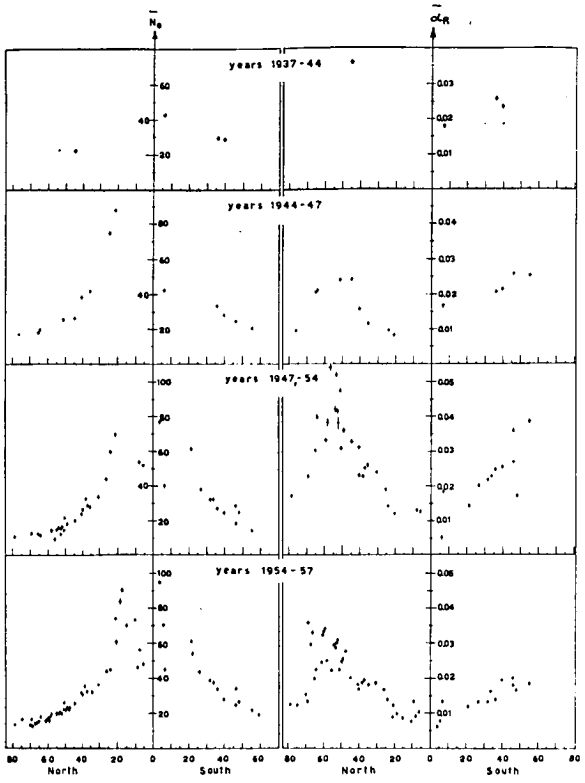


FIGURE 1.—Correlation of the long-term variation $\bar{N}$ and $\bar{R}$. The abscissae, in this figure and in the following (figures 2, 3, 5, 6, 7 and 8) is the "effective" latitude as defined in section 3.1.

the years 1938–1944. Comparison of results for Northern and Southern Hemispheres indicated a noticeable asymmetry in the correlations with A_F, and in any case a clear latitudinal effect with a minimum at the equator.

In order to improve the above results and to see if they also apply to other phases of the solar cycle, we have since investigated the correlations of electron densities with as many solar parameters as possible. First we calculated the simple correlations of the electron densities with the sunspot number R, the areas A_R, A_F, $A_{\Phi H}$ and $A_{\Phi Ca}$; later we looked for possible effects of the position of the perturbation centers on the solar disk by investigating some correlations with the Meudon filament activity.

The overall results of such calculations give a conclusive confirmation of a latitude effect of the regression coefficients in the four intervals 1938–1944, 1944–1947, 1947–1954, 1954–1957 which correspond alternately to decreasing and increasing phases of the solar cycle. For a few observatories we could obtain only qualitative rather than quantitative information, because of insufficient data. These cases are not included in the quantitative results we given in the next sections.

We consider separately the cases of the long-term and month-to-month variations.

3.1 The Case of Long-Term Variation $\bar{N}$

The principal results are given in figures 1, 2, and 3, in which we show the quantities $\bar{N}_0$ and $\bar{\alpha}_i$ defined by (1) for the simple correlations of $\bar{N}_0$ with the sunspot number $\bar{R}$ (fig. 1), the area $\bar{A}_F$ (fig. 2) and the areas $\bar{A}_{\Phi H}$ and $\bar{A}_{\Phi Ca}$ (fig. 3),

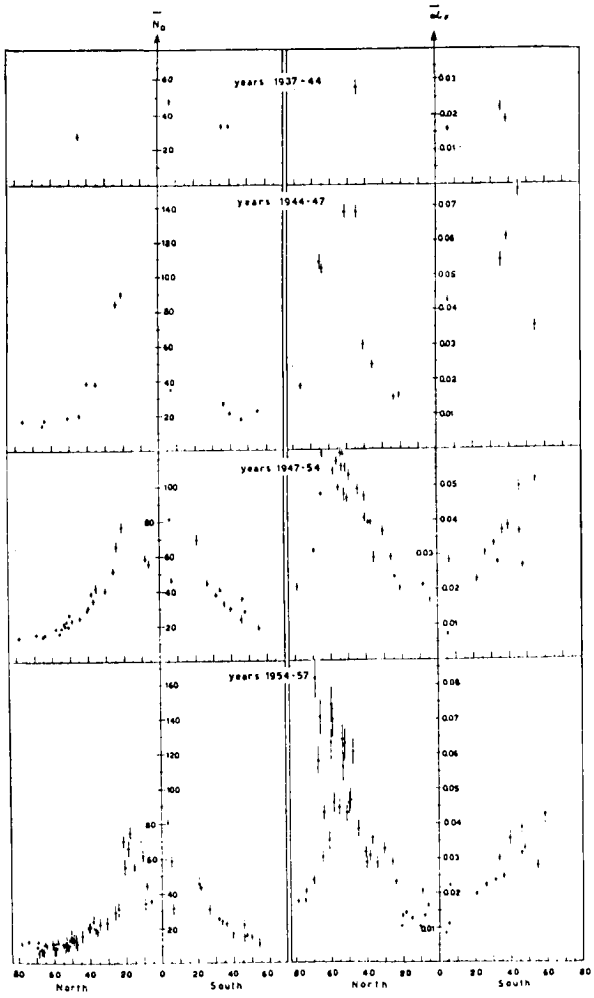


FIGURE 2.—Correlation of the long-term variation in $\bar{N}$ and $\bar{A}_F$.

respectively. In all the diagrams we indicate on the abscissae an "effective" latitude defined as the arithmetical mean of geographical and geomagnetic latitudes. As found earlier in M1, such a choice leads to some reduction of minor irregularities in the distribution of the points on the graphs and takes into some account the dependence of the electron density upon both the geographical and geomagnetic coordinates. In the left half of each figure we show the values of $\bar{N}_0$, in the right half the values of $\bar{\alpha}_i$. The amplitude of the standard errors is indicated for all the calculated points.

A first interesting feature is the rather symmetrical distribution of $\bar{N}_0$ in the two hemispheres for latitudes greater than 20°. In the tropical latitude belt $\bar{N}_0$ exhibits two relative maxima; the maximum in the Northern Hemisphere is rather localized while the position of the Southern Hemisphere maximum is not as clear, because of the lack of sufficient experimental data at those latitudes. The occurrence of such maxima in tropical regions is a well known feature of the F2 layer.

We can also compare the latitudinal variation of $\bar{N}_0$ and the annual mean of $\cos\chi_{\text{noon}}$, where χ_{noon} is the noon zenith distance of the sun. Values of $\cos\chi_{\text{noon}}$ are proportional to the noon ionization intensity and, for the F2 layer, also

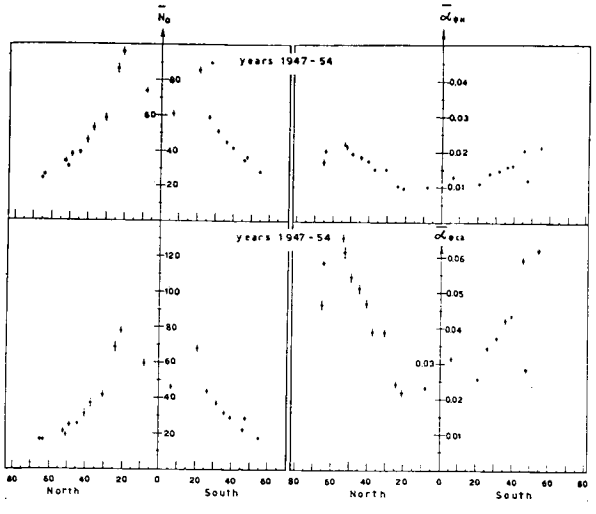


FIGURE 3.—Correlation of the long-term variation in $\bar{N}$ and $\bar{A}\phi_{\text{H}}$ (upper graphs) and $\bar{N}$ and $\bar{A}\phi_{\text{C}}°$ (lower graphs).

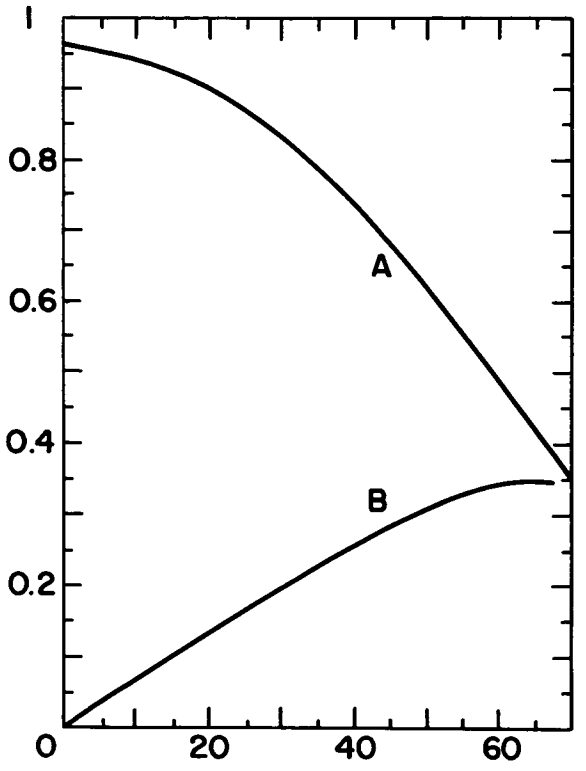


FIGURE 4.—Latitudinal variation of annual mean of $\cos\chi_{\text{noon}}$ (curve A) and of the amplitude of its seasonal variation (curve B). The abscissa is the geographical latitude.

approximately proportional to the noon electron density. These values have been calculated (fig. 4), taking into consideration the sphericity of the earth and sunlight at ionospheric levels (even during polar winter at noon in all the northern observatories). By use of proper normalization factors, we obtain a rather good fit of the experimental values $\bar{N}_0$ with the annual mean of $\cos\chi_{\text{noon}}$ for latitudes higher than 30°–35°.

Some anomalous values are present in the latitudinal distribution of $\bar{N}_0$, in particular the rather high values at Huancayo and Leopoldville, which are at or very close to the geomagnetic equator. Generally, as mentioned earlier, local conditions (such as upper atmospheric winds, anomalous temperature or concentration distribution) can influence the electron density. At this stage, then, we try to give a physical meaning to the general behavior rather than to some particular anomaly, which must await a more refined

theory. In the particular case of equatorial data, for example, we recall that the geomagnetic field also exhibits an anomalous behavior on the geomagnetic equator, because of the equatorial electrojet at ionospheric levels.

We conclude that the values $\bar{N}_0$, for latitudes higher than 30°–35° are a rather good index of the steady ultraviolet radiation at F2 levels.

If we now take into account the latitudinal variation of the regression coefficients $\bar{\alpha}_i$, assuming a dependence of $\bar{N}$ on the solar activity as given in (1), we would expect, within the statistical and experimental errors, constant values at all latitudes. This is not the case, however; a remarkable latitudinal effect is apparent.

A first interesting feature appearing in this latitudinal variation of the coefficients $\bar{\alpha}_i$ is an appreciable difference between their values in the Northern and Southern Hemispheres, when we consider the correlation of $\bar{N}$ with $\bar{R}$ or $\bar{A}_F$. However, this is not the case for correlations with $\bar{A}_{\Phi H}$ or $\bar{A}_{\Phi Ca}$. The existence of a difference between the two hemispheres was already shown in M1 for the years 1947–1954. However, it may be only apparent, due to the particular solar parameters used in M1. Actually, the areas $\bar{A}_{\Phi H}$ and $\bar{A}_{\Phi Ca}$ may be a better index of solar activity than are the parameters $\bar{R}$ and $\bar{A}_F$ used in M1.

The most surprising feature of the latitudinal effect is the sharp maximum of $\bar{\alpha}_i$ at latitudes of 55°–60°. At first sight, an increase of $\bar{\alpha}_i$ with latitude could be due to the decreasing values of $\bar{N}_0$. Such an effect can to some extent affect the variation of $\bar{\alpha}_i$. However it does not seem to be the determining factor, since at northern latitudes higher than 60°, $\bar{\alpha}_i$ first decreases remarkably and very steeply and then recovers to the equatorial values. For extreme southern latitudes the decrease of $\bar{\alpha}_i$ cannot be checked because of the lack of data; only its increase to 55° is observed.

A simple physical explanation of the above latitudinal effect can be given in terms of two superimposed effects. The first is a variation of ultraviolet radiation within the solar cycle, which is present in all places. The other effect is a latitudinal variation in corpuscular radiation directly or indirectly coming from, or influenced by, the sun. If this is the case, one can study the latitudinal variation of the absolute "corpuscular" effect.

We can write equation (1) in the form

$$\bar{N}=\bar{N}_0[1+a_i\bar{A}_i]+\bar{N}_0(\bar{\alpha}_i-a_i)\,\bar{A}_i, \qquad (5)$$

where the first term represents the "ultraviolet" effect and the last one the "corpuscular" effect; the quantity a_i is assumed to be independent of latitude. A possible value of a_i may be the mean of the $\bar{\alpha}_i$ values for tropical latitudes, which is practically the same as that for very high latitudes.

A clear latitudinal effect is again apparent not only in $\bar{\alpha}_i-a_i$, but, above all, in $\bar{N}_0(\bar{\alpha}_i-a_i)$. This last feature gives further evidence of a latitude dependence of the absolute intensity of the "corpuscular" effect. For reasons of brevity we do not give diagrams of $\bar{N}_0(\bar{\alpha}_i-a_i)$; we will rather consider the correlations of month-to-month variation which we study in the next section and which, from the physical point of view, are much more significant.

It should be emphasized that the $\bar{\alpha}_i$ distribution has its maximum at lower latitudes than the auroral zone; the lines of force between the geomagnetic latitudes 30° and 65° cross the equatorial plane at geocentric distances of about 1.4 and 5.5 earth's radii, i.e. at heights of 2500 and 29,000 km above the ground. At least from a qualitative point of view, one could think in terms of some connection between corpuscular ionospheric effects and Van Allen radiation belts. We shall consider this point later in section 4.

The double correlation with $\bar{R}$ and $\bar{A}_F$, defined by (2) was also tested, for the years 1944–1947 and 1954–1957; the same double correlation with $\bar{R}$ and $\bar{A}_F$, for the years 1947–1954, having been already considered in M1. For the latter time interval a latitudinal constancy of $\bar{\beta}_R$ and a latitudinal variation of $\bar{\gamma}_F$ was established; such a feature is not confirmed during the two increasing phases of the solar cycle, so that it must be considered as a nonpermanent feature.

3.2 The Case of the Month-to-Month Variation $N-N_{12}$

This case is more expressive from the physical point of view, because it takes into account the actual unsmoothed month-to-month variation of both electron density and solar parameters.

The results of our calculations are shown in figures 5, 6, and 7. In this case, obviously, the statistical errors are about an order of magnitude greater than in the case of the long-term variation. The essential features of the behavior of $(N-N_{12})_0$ and α_i are clear. The statistical improvement with a sufficient amount of ionospheric data can be seen by considering the 11-year period 1947–1957. The mean values of α_i for correlations with monthly values of R and A_F are again greater in the Northern than in the Southern Hemisphere, while this is not the case for the correlations with $A_{\Phi H}$ and $A_{\Phi Ca}$ (fig. 7).

If we again assume a mean value a_i of α_i for the tropical belt of latitudes, we can write the regression equation (3) in the form

$$N-N_{12}=(N-N_{12})_0[1+a_iA_i+(\alpha_i-a_i)A_i]. \quad (6)$$

Figure 8 shows the 1947–1957 behavior of

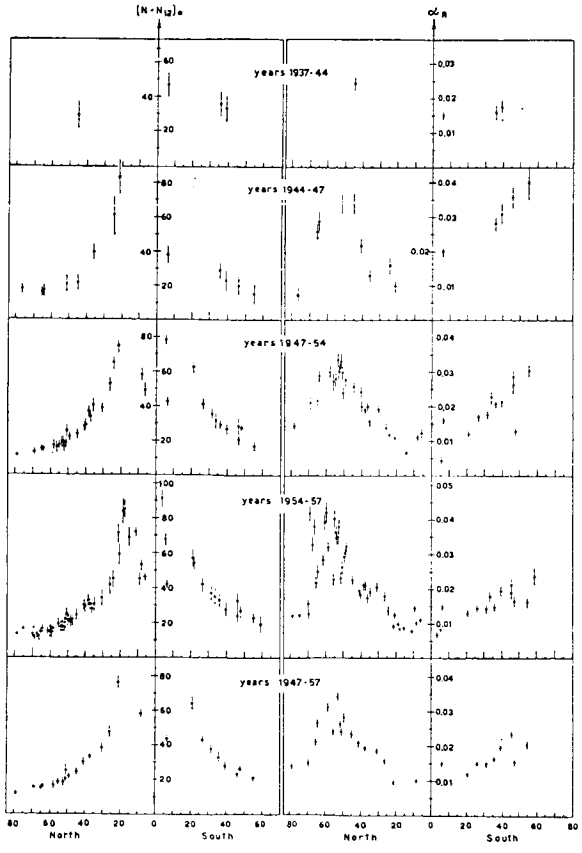


FIGURE 5.—Correlation of month-to-month variation N–N_{12} and R.

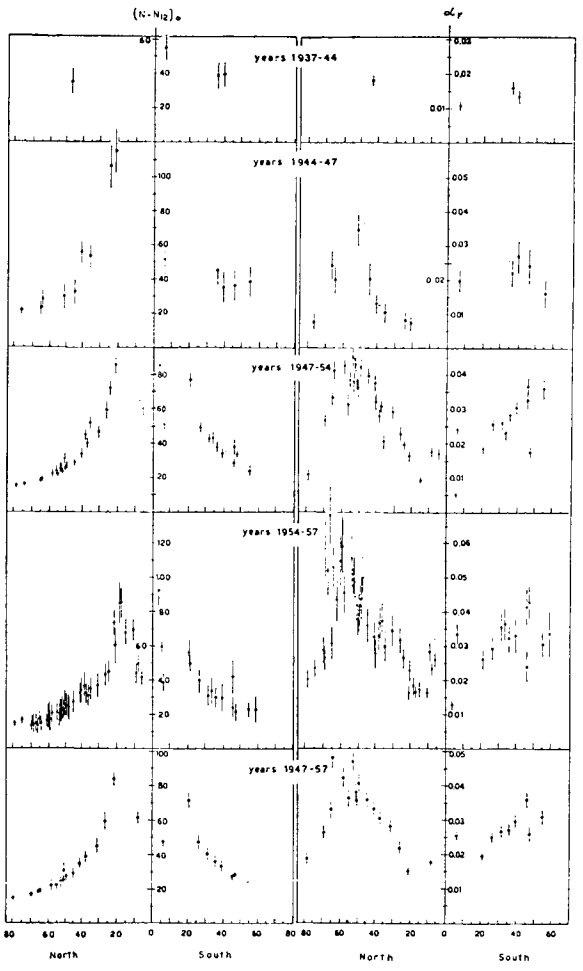


FIGURE 6.—Correlation of month-to-month variation N–N_{12} and A_F.

$(N-N_{12})_0\times(\alpha_i-a_i)$ which represents, as we have already seen, an index of the absolute latitudinal effect on the ionospheric electron density.

Although the statistical errors are much greater because of the combination of the errors in $(N-N_{12})_0$ with those in α_i and a_i, the latitudinal effect is clearly enhanced.

For the double correlation defined by (4), however, we again failed to obtain results having some clear physical meaning.

3.3 Effect of the Heliographic Latitude of Solar Disturbances

In order to look for a dependence of ionospheric electron density on the heliographic lati-

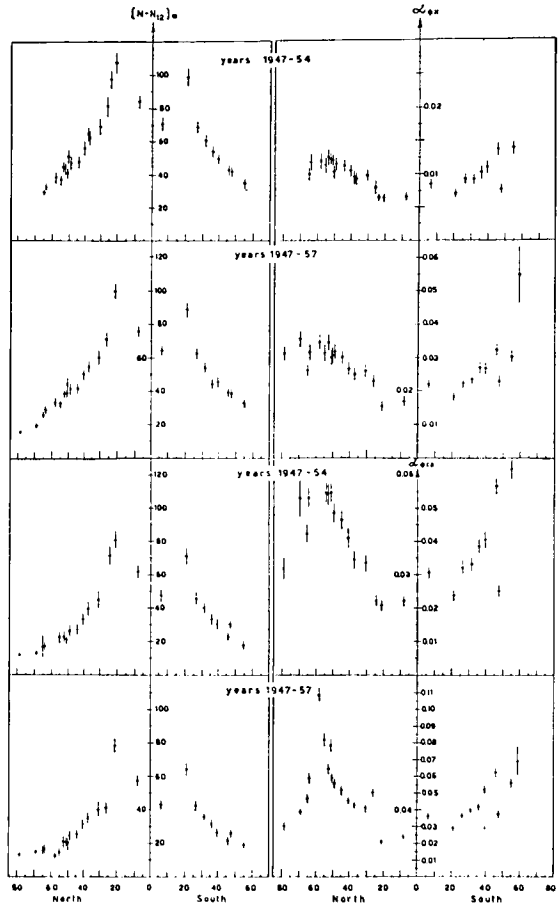


FIGURE 7.—Correlation of month-to-month variation $N-N_{12}$ and $A\phi_H$ (upper graphs) and $N-N_{12}$ and $A\phi_{Ca}$ (lower graphs).

tude of the perturbation centers on the solar disk, we studied the data collected by the observatory of Meudon on filament activity, as a function of heliographic coordinates. Such data are given for each solar rotation period (27 days). We expressed them in terms of terrestrial months by making a weighted mean as follows: if in a terrestrial month there are contained n days of the 27 days of a solar rotation period and m days of the following rotation period, and x and y are the numbers which express the filament activity for the two rotation periods, the mean terrestrial-month activity is defined as $(nx+my)/(n+m)$.

We have separately considered the activity of: (i) polar filaments; (ii) equatorial filaments; (iii) filaments between the heliographic latitudes 30N to 30S; (iv) filaments in the Northern Solar Hemisphere; (v) filaments in the Southern Solar

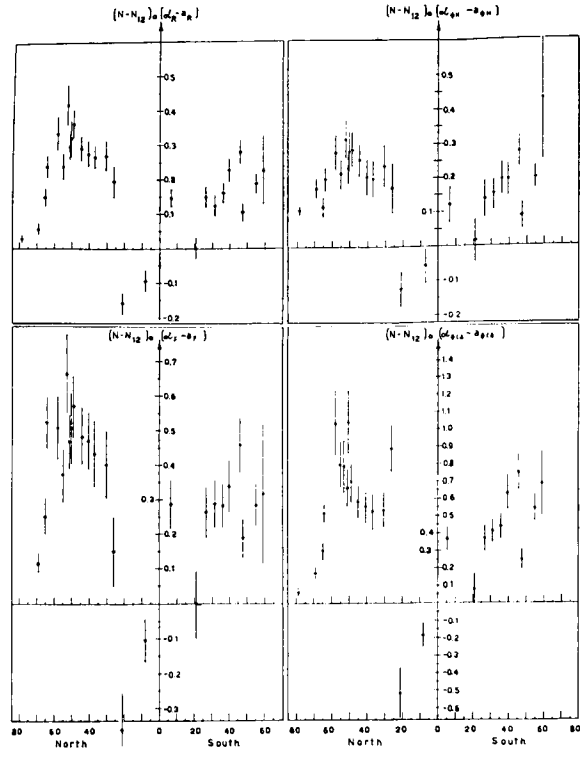


FIGURE 8.—Latitudinal variation of $(N-N_{12})_0(\alpha_i-a_i)$ for the 11 years 1947–57. The values at latitude 59S refer to the years 1954–57.

Hemisphere. In each case, the correlations do not seem significant; the values of the coefficients of the linear regressions (1) and (3) exhibit noticeable and irregular differences for different observatories; in many cases negative values of $\bar{N}_0$ and $(N-N_{12})_0$ were also obtained.

According to the present investigation, no significant correlation appears to exist between the ionospheric F2 layer and the latitudinal distribution of solar activity centers.

4. THE INTERPRETATION OF THE LATITUDINAL EFFECT

4.1 The "Corpuscular" Hypothesis

On the basis of the conclusion drawn from the results of section 3, we give a tentative physical interpretation of the latitudinal effect according to the following hypotheses.

(i) The effect is caused by a corpuscular radiation impinging, with a characteristic latitudinal distribution, upon the upper atmosphere.

(ii) The corpuscular radiation has a rather low mean energy (of the order of Kev or tens of Kev), so that it reaches the 200-km level, i.e., the F2 layer, but does not necessarily reach the E layer (Mariani, 1957).

(iii) The intensity of the radiation depends (as a first approximation) linearly on the solar activity.

First, we estimate the order of magnitude of the particle flux necessary to give the actual ionization rate. We assume that a particle loses all its energy in the ionization of neutral particles. In the steady state which we may assume for the F2 layer at noon, we write the equilibrium equation in the form

$$I-bn_e=0, \tag{7}$$

where I is the production rate, b is the effective recombination coefficient, and n_e the electron density. By noting that the maximum values of the ratios $(\alpha_i-a_i)/a_i$ at latitudes of 55°–60° generally have values between 1 and 2, one concludes that the effects of the variable part of the ultraviolet radiation and that of the corpuscular radiation are comparable. However, the total ultraviolet contribution, proportional to $1+\alpha_i A_i$, becomes more and more important with respect to the corpuscular contribution as the solar activity decreases. At maximum solar activity (a conventional value of the sunspot number R equal to 150) the quantities $1+\alpha_i A_i$ and α_i-a_i are about the same; in this physical situation, the corpuscular ionization rate I_c is comparable to the ionization rate I_u of the ultraviolet radiation.

Thus we have

$$I_u\approx I_c\approx\tfrac{1}{2}I\approx b\tfrac{1}{2}(n_e). \tag{8}$$

If we put $b=10^{-4}$ sec^{-1} and $n_e=10^6$ cm^{-3} (so that the ionospheric layer has a critical frequency $f_c\approx 9$ Mc sec^{-1}), we obtain, at the height of maximum electron density (300–350 km):

$$I_u\approx I_c\approx 50 \text{ ions cm}^{-3}\text{ sec}^{-1}.$$

We now make use of the fact that the photoionization rate at heights more than 300 km decreases exponentially (Watanabe and Hinteregger, 1962). We assume that above this altitude 50 percent of the ionization is due to corpuscular radiation. Thus an estimate of the total corpuscular ionization in a square cm column of air above 300 km at middle latitudes can be $I_{\text{total}}\approx 2\times10^9$ ions cm^{-2} sec^{-1}. If the mean ionization loss is 35 ev/ion-pair, the corresponding kinetic energy flux is $F_T\approx 7\times10^{10}$ ev cm^{-2} sec$^{-1}\approx 0.1$ erg cm^{-2} sec^{-1}. For comparison, the total energetic flux entering the atmosphere, assuming the solar constant = 2 cal cm^{-2} min^{-1} is 1.4×10^6 erg cm^{-2} sec^{-1}.

The above data allow a simple evaluation of the order of magnitude of the required particle flux. If the particles can penetrate vertically to a minimum altitude of 150 km, they have, if they are electrons, a kinetic energy $T\approx 1500$ ev. However, the effect of scattering and of the geomagnetic field is that the path of the electrons is not a vertical straight line. On the average, then, they stop at higher altitudes, i.e., in the F layer. The total incoming electron flux F_e, at maximum solar activity, can be estimated as

$$F_e=F_T/1500\approx 5\times10^7 \text{ electrons cm}^{-2}\text{ sec}^{-1}.$$

The velocity of an electron having an energy $T\approx 1500$ ev is $v\approx 2.3\times10^9$ cm sec^{-1} so that the particle density in the incoming stream is of the order of 10^{-2} cm^{-3}.

If such a corpuscular radiation impinges uniformly on the upper atmosphere between the latitudes of 30° and 65° (a surface of one third of the total area of the earth), the total incoming flux is of the order of 10^{26} electrons sec^{-1}. In view of this intensity, it does not seem possible to attribute such a flux of electrons to a primary source other than the sun. On the other hand, the particular features of the latitudinal effect exclude the possibility that they are coming directly from the sun along Störmer trajectories. As mentioned earlier, we may however assume that the electrons are leaking from the Van Allen radiation belts.

We may also consider the low-energy electron fluxes measured by Krasovskii et al. (1962) and, more recently, by O'Brien (1962). Krasovskii reports that at 45° geomagnetic latitude an energy flux of dumped electrons of energy $T\approx 10$ Kev between 10^{-2} and 1 erg cm^{-2} sec^{-1} was observed during one pass of Sputnik 3. The more systematic measurements of O'Brien, made on Injun I, gave median fluxes of 10^6 trapped electrons cm^{-2} sec^{-1} with $T\geq 40$ Kev. Corresponding to these intensities, O'Brien reports average energy fluxes

of trapped electrons with $T \geq 1$ Kev of the order of 1 to 10 erg cm^{-2} sec^{-1}. We notice that the above assumed electron energy of 1.5 Kev is just the average energy if we assume, for energy ≥ 1 Kev, a power law differential spectrum with an exponent $\gamma = 4$. Thus, the number fluxes of trapped electrons with mean energy $T = 1.5$ Kev are of the order of 5×10^9 cm^{-2} sec^{-1}. With "lifetimes" of 10^3 to 10^4 sec, as calculated by O'Brien for particles with $T > 40$ Kev, one computes dumped electron fluxes of not less than 10^5 to 10^6 cm^{-2} sec^{-1}; the corresponding energy fluxes are not less than 10^{-4} to 10^{-3} erg cm^{-2} sec^{-1}.

If one considers that the measurements of O'Brien have been made in a period of reduced solar activity ($R = 50$) and that the "lifetime" of very low-energy electrons may be substantially lower than that estimated for electrons with $T > 40$ Kev, we can effectively consider the dumped electron flux from Van Allen belts as an important source (or *the* source) of the corpuscular radiation in the F2 layer.

Some difficulty could arise here from the fact that the corpuscular ionospheric effect decreases very sharply at latitudes higher than 60°, while in practice the electron fluxes measured by O'Brien are nearly constant at all latitudes up to 70°, corresponding to a maximum geocentric distance of about 10 earth radii. Concerning this point, we must bear in mind, however, that our results concern the state of the ionosphere averaged with respect to the solar cycle. One cannot exclude some small year to year shift of the latitude of maximum corpuscular effect, in particular toward lower latitudes as solar activity decreases. If this is the case, one can reasonably assume that at maximum solar activity the maximum corpuscular effect may occur at higher geomagnetic latitudes, 60° to 65° or even higher.

A further quantitative element we can easily calculate is the integrated flux of particles during the entire 11-year solar cycle. If we assume a linear long-term time variation, and remember that the above calculated flux values refer to the maximum solar activity, we estimate that the integrated electron flux entering the upper atmosphere may be 10^{34} electrons. The corresponding integrated energy, assuming a mean energy of Kevs, is 10^{25} to 10^{26} ergs.

One immediately sees that the above integrated fluxes of particles and energy represent only a very small part of the fluxes ultimately emitted from the sun during its full cycle of activity.

4.2 Other Evidence for Corpuscular Effects and Conclusions

Finally, we want to draw attention to the many experimental indications of intensive particle fluxes in the upper atmosphere. It is agreed that polar aurorae are the result of corpuscular radiation, although the origin is not fully understood.

Antonova and Ivanov-Kholodny (1961) point out the possibility of a corpuscular origin of nighttime ionospheric ionization; their calculated flux of electrons of about 100 ev would be 10^{10} to 10^{11} electrons cm^{-2} sec^{-1}, to which corresponds an energy flux of 1 to 10 erg cm^{-2} sec^{-1}. These electron and energy fluxes are 10^3 to 10^4 higher than our minimum estimates. Such high fluxes, based on the assumption of a high value of the loss coefficient in the F2 layer (10^{-7} to 10^{-6} cm^3 sec^{-1}) are probably overestimated, since the actual coefficient appears to be much smaller.

Several experimental results (Boudeau and Bauer, 1962) indicate that, at altitudes approximately between 150 and 350 km, the electron temperature is higher than the ion temperature. Bourdeau (1962) also points out that "large fluxes of quasi-energetic particles which could provide an additional ionization source have been observed at some geographical locations in the upper atmosphere."

Harris and Priester (1962), in their calculated theoretical models for the solar-cycle variation of the upper atmosphere, assume an ultraviolet heat source and a corpuscular heat source, of almost equal magnitude, of the order of 1 erg cm^{-2} sec^{-1}. Contrary to our scheme, however, the assumed ratio between corpuscular and ultraviolet fluxes does not vary during a solar cycle.

We conclude that if the electron flux calculated in section 4.1 is really present in the F2 layer, it could possibly originate in the radiation belts; in any case the Van Allen belts are a good reservoir of ionospheric ionizing particles. This possibility obviously does not exclude some other acceleration mechanism acting on very low-energy electrons "normally" present in the upper atmosphere:

for example, one could think in terms of an electric field, present only (or mainly) in the latitude range in which the "corpuscular effect" is present. From a general point of view, it will be very interesting to study the eventual dependence of the corpuscular effect on local time, as has been done for the polar aurora. Actually, the experimental parameter, the F2 layer critical frequency, is a more or less approximate index of ultraviolet and corpuscular radiation only near noon; at other times, particularly during the night, it is controlled primarily by other phenomena as time and height variations of recombination or attachment coefficients, temperature variations, convective motions, etc., so that it cannot be considered an even approximative index of the incoming corpuscular radiation. At present, the first and more immediate experimental problem would be the *direct* detection of very low-energy particle flux at ionospheric levels; we hope that this can be done in the near future.

ACKNOWLEDGMENTS

I take this opportunity of thanking the many people who have sent me the large volume of ionospheric data used in this study; in particular, Mr. W. B. Chadwick of the National Bureau of Standards, Boulder, Colo. A great quantity of data has been analyzed by the electronic computer of the "Centro di Studi sulle Calcolatrici Elettroniche" of the University of Pisa. I express my sincere thanks to Professor M. Conversi, director of the "Centro," and to Dr. E. Fabri, whose help in preparation of the computer programs was indeed very valuable.

REFERENCES

ANTONOVA, L. A., and IVANOV-KHOLODNY, G. S., Ionization in the night atmosphere (corpuscular hypothesis), *Proc. Second Internat. Space Sci. Symposium*, Florence, 981–992, 1961.

BOURDEAU, R. E., *Space flight studies of the ionosphere.* Preprint NASA University Conference on Space Sciences and Technology at Chicago, NASA publication X–615–62–204, 1962.

———, and BAUER, S. J., Structure of the upper atmosphere deduced from charged particle measurements on rockets and the Explorer VIII satellite, *Proc. Third Internat. Space Sci. Symposium*, Washington, 173–193, 1962.

O'BRIEN, J., Lifetimes of outer-zone electrons and their precipitation into the atmosphere, *J. geophys. Res.*, **67,** 3687–3706, 1962.

HARRIS, I., and PRIESTER, W., Theoretical models of the solar-cycle variation of the upper atmosphere, *J. geophys. Res.*, **67,** 4585–4592, 1962.

KRASOVSKII, V. I., SHKLOVSKII, I. S., GAL'PERIN, YU. I., SVETLISKII, E. M., KUSHNIR, YU. M., and BORDOVSKII, G. A., Detection of electrons with energies of approximately 10 Kev in the upper atmosphere, *Planetary Space Sci.*, **9,** 27–40, 1962.

MARIANI, F., Sulle Correlazioni tra Densita Elettronica Ionosferica e Attivita Solare, *Ann. Geofis.*, **10,** 71–87, 1957.

———, The worldwide distribution of the F2 layer electron density: seasonal and non-seasonal variations and correlations with solar activity, *Nuovo Cimento*, **12,** 218–240, 1959.

———, Correlation of F2 layer electron density and solar activity in the years 1938–1944, *Proc. of Liege Symposium on Electromagnetic Wave Propagation*, October 1958, London and New York, Academic Press, 451–457, 1960.

WATANABE, K., and HINTEREGGER, H. E., Photoionization rates in the E and F regions, *J. geophys. Res.*, **67,** 999–1006, 1962.

PITCH ANGLE DISTRIBUTION OF THE PHOTOELECTRONS AND ORIGIN OF THE GEOMAGNETIC ANOMALY IN THE F2 LAYER

F. MARIANI*

Goddard Space Flight Center

A major anomaly of the F region is the "equatorial" or "geomagnetic" anomaly, which consists of the pronounced minimum of f_0F2 on the geomagnetic equator and two maxima near 20° north and south geomagnetic latitude, at meridian hours and in the afternoon, in every season. A recent theoretical approach (Goldberg and Schmerling, 1963) favors the hypothesis that a diffusion process from the equator along the magnetic lines of force can produce the enhancement of f_0F2 in both hemispheres. However, it has been suggested that this process alone is not sufficient to explain completely the observed anomaly (Rishbeth, Lyon and Peart, 1963).

It is the purpose of this letter to present in some detail one simple mechanism which also may play a role in the explanation of the anomaly. This mechanism is based on a more detailed analysis of the physical processes that the photoelectrons undergo in the upper atmosphere.

It is generally assumed that the photoelectrons produced in the atmosphere by solar ultraviolet radiation undergo physical processes which contribute to the F region electron distribution only at the altitudes where they are produced. In this hypothesis no importance is attached to the actual angular distribution of the photoelectrons produced by the photoelectric effect. This distribution (Heitler, 1944) is peaked at right angles with respect to the direction of the incident photons, according to a $\sin^2\theta$ law (θ is the angle between the photoelectron velocity vector and the incident photon direction). On the other hand, the incident photon beam, as a whole, is not polarized, so that the azimuthal photoelectron distribution is on the average practically uniform.

At rather low altitudes, in the E and F1 layers, the thermalization lifetime of the produced photoelectrons is very short, due to their small mean free path with respect to all the possible physical loss-processes. However this is not the case for the upper F2 region, where we may assume (Hanson, 1962) that the most important energy loss of the photoelectrons occurs in elastic collisions with ambient electrons. An immediate consequence of the large mean free path is that the photoelectrons are subject to some motion after their production rate can differ from the case of "local" thermalization. The perturbation is related to latitude and to the time-variable effect of the geomagnetic field on the $\sin^2\theta$ angular distribution of the ejected photoelectrons. We suggest that what is called "anomaly" could just be, at least partially, the result of the redistribution of the photoelectrons produced locally, due to the presence of the geomagnetic field. Anyway, we must point out that the experimental datum we have is the density of the ambient thermal electrons, so that its comparison with the photoproduction rate remains to some extent arbitrary.

For a centered magnetic dipole with its axis aligned with the geographical axis, the corresponding pitch angle distribution is given by the function (Mariani, 1963)

$$g(\alpha)=\tfrac{3}{4}\sin\alpha[\sin^2\alpha+\sin^2\beta(1-\tfrac{3}{2}\sin^2\alpha)]$$

The quantity $g(\alpha)\,d\alpha$ gives the fraction of electrons$\cdot$cm$^{-3}\cdot$sec^{-1} with pitch angle α between α

*NAS–NASA Senior Postdoctoral Research Associate, on leave of absence from the University of Rome.

and $\alpha+d\alpha$. The angle β is the angle between the direction of the incident photon and the geomagnetic field. This angle β is a function of the geographical latitude Λ, the magnetic inclination I, the local time Φ and the solar declination δ according to the expression

$$\cos\beta = -\sin\delta\cos(I+\Lambda)+\cos\delta\sin(I+\Lambda)\cos\Phi$$

In figure 1 the function $g(\alpha)$ is plotted as a function of pitch angle for different values of β. The photoelectrons with rather small pitch angles can drift along the magnetic line of force within a narrow cylinder whose radius has an upper limit of some tens of centimeters. The photoelectrons with high pitch angles, instead, cannot migrate significantly. The fraction of photoelectrons with intermediate pitch angles α actually able to drift along the magnetic lines of force increases sharply when β increases.

Figure 2 shows the percentage of photoelectrons $\cdot$ cm$^{-3}\cdot$ sec^{-1} having pitch angles between 0 and a maximum angle $\alpha_m=54.6°$ for different times at the equinoxes and the solstices. We consider now what happens at low latitudes. The

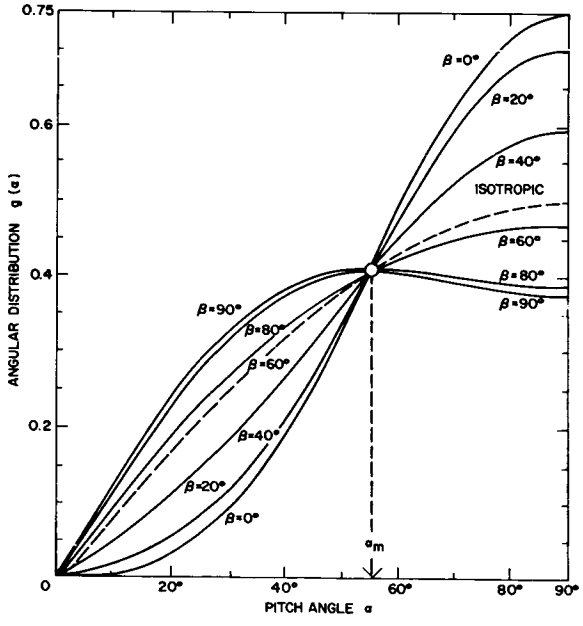


FIGURE 1.—Pitch angle distribution of the photoelectrons for different values of the angle β. The dotted line gives the distribution in the case of an isotropic angular distribution. The angle α_m is the abscissa of a point common to all curves.

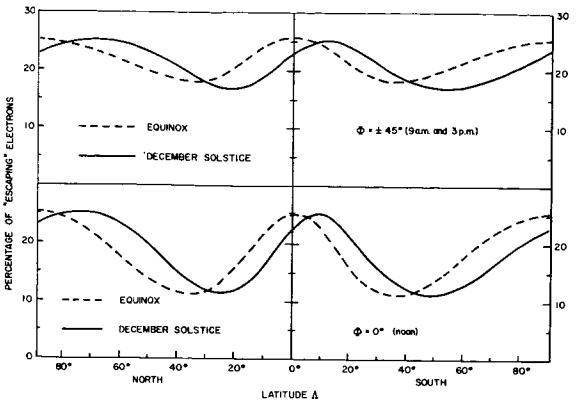


FIGURE 2.—Percentage of "escaping" photoelectrons. cm$^{-3}\cdot$sec^{-1} (electrons with pitch angle less than α_m) as a function of latitude, at different times and seasons. The local time is expressed by the hour angle ϕ ($\phi=0°$ at noon).

less energetic of the upgoing photoelectrons are gradually stopped along their spiral path toward the equatorial plane and only the more energetic ones can possibly reach the conjugate point. The photoelectrons going down are instead lost in the lower ionosphere where stopping processes other than elastic collisions with ambient electrons play a predominant role. Thus, their relative contribution to the existing photoelectron production rate is obviously small; this is not the case of the upgoing photoelectrons, at heights where the local production is very small.

Our problem is to estimate the "actual" production at a given place A_i (fig. 3). In the semiqualitative approach given in this letter, we make the following assumptions:

(a) All photoelectrons produced below the 300-km level cannot escape, independent of their energy.

(b) The photoelectrons produced above 300 km can escape according to the pitch angle distribution $g(\alpha)$ only if $\alpha\leq\alpha_m=54.6°$.

(c) The escaping photoelectrons lose their energy in elastic collisions with ambient electrons without undergoing, on the average, appreciable variation of their pitch angle α.

We remark that, at least at low latitudes, the height profile of the photoproduction rate above the 300-km level at noon is almost insensitive to

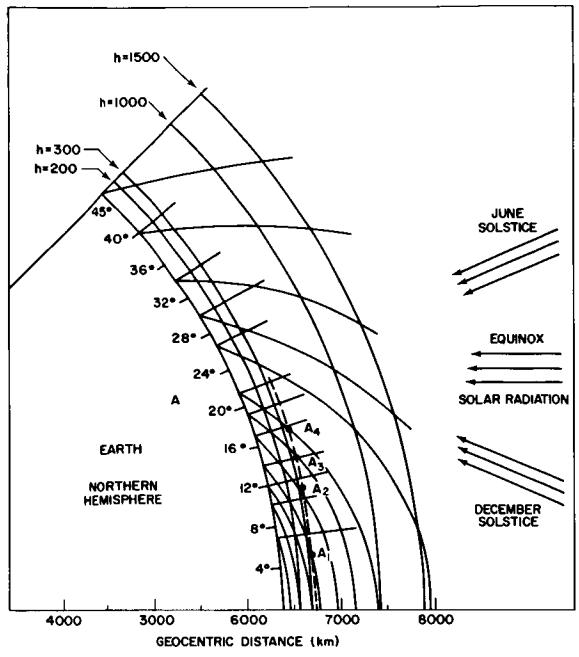


FIGURE 3.—A representation of the geomagnetic lines of force. Different constant geometrical height lines are shown, too. The dotted region indicates the height interval in the F2 layer, within which all the energetic photoelectrons are "locally" lost. The direction of solar incoming radiation at equinoxes and solstices is indicated.

latitude variations, so that we can make the further hypothesis:

(d) The production rate between 40°N and 40°S above 300 km is a known function of the height h, *only*.

At the points A_1, A_2, A_3, etc., at a constant height h above the ground, the fraction of photoelectrons able to escape upward or downward along the magnetic lines of force is steadily decreasing from 0° to 35°, at the equinoxes. As a consequence of this photoelectron drift, the "actual" production is appreciably greater at tropical latitudes than at the equator.

On the other hand we must also take into account the fact that, due to the very small ambient electron density at heights above 1000 km, a certain latitude-increasing part of the escaping photoelectrons can possibly penetrate, more or less deeply according to their energy, the upper ionosphere into the opposite hemisphere. This effect can balance part of the "magnetic depletion" effect at latitudes above about 20°.

In practice, the largest absolute contribution to the escaping photoelectrons is confined to the height interval between 300 and 500 km. The total number of escaping photoelectrons has been estimated (Hanson, 1962) to be of the order of 10^8 $cm^{-2}sec^{-1}$. If they were uniformly lost along a line of force they could given an average supplementary "production" density of 10^8/(average length of magnetic line), whose upper limit cannot be more than 1 electron·cm^{-3}·sec^{-1}. Such a "production" is possibly important only in the equatorial region where it is distributed at higher levels at other latitudes. The net effect could give some contribution to the experimentally observed higher altitude of maximum electron density at the equator, despite the lower altitude of the maximum of the photoelectron production.

The magnetic depletion effect increases again at latitudes above approximately 35°. If we take into consideration the latitudinal variation of the atmospheric structure, which we completely neglected here, and the minor effect between conjugate points, it seems reasonable to expect some maximum of ionospheric electron density at some intermediate latitude, possibly just in the tropical region. On the other hand, we must expect some appreciable longitudinal dependence of the magnetic depletion effect due to actual inclination of the magnetic dipole axis with respect to the geographic axis.

Concerning the contribution of the magnetic depletion at other times than noon, we see (fig. 2) that the effect is maximum just at noon; no differential latitude depletion occurs at the equinoxes at 6 a.m. (or 6 p.m.), because the angle β is zero. However, the geomagnetic anomaly in the later afternoon or after sunset is obviously an effect of the hysteresis (or, in other words, of the long lifetime) of the F2 layer.

The situation in solstice-months, for example in December, may be as follows: We see, from figure 2, that in this case the magnetic depletion effect is a maximum at 10° latitude south and the minimum depletion occurs when $A \sim 25°$ in the northern (or winter) hemisphere and $A \sim 50°$ in the southern (or summer) hemisphere. There must be, in other words, some southward shift of

the equatorial minimum of the photoelectron production. This is just what appears to be the case, according to the analysis by Thomas (1963).

Moreover, within the limitations given by the hypothesis (d), we can also expect a higher electron production in the winter hemisphere at low latitudes. This could result in an important contribution to the winter anomaly in this latitude range.

At higher latitudes the percentage magnetic depletion effect is greater in the winter hemisphere than in the summer one. This is just in the opposite sense to contribute to the winter anomaly. However no definite conclusion can be drawn in this latitude range without taking into proper account the latitudinal variation of the atmospheric structure which is certainty much more effective at solstices than at equinoxes in giving latitudinal variations and differences between northern and summer hemispheres.

A more quantitative approach to the problem encounters difficulties because one must know, at least approximately, the energy spectrum of the photoelectrons and the true-height profile of ambient electrons over a rather extended range of latitudes and times. Work along these lines is now in progress.

ACKNOWLEDGMENTS

I am very indebted to Dr. S. J. Bauer for helpful comments and discussions.

REFERENCES

GOLDBERG, R. A., and SCHMERLING, E. R., "The Effect of Diffusion on the Equilibrium Electron Density Distribution in the F Region near the Magnetic Equator," *J. Geophys. Res.*, **68**, 1927–1936, 1963.

HANSON, W. B., "Electron Temperatures in the Upper Atmosphere," *Proc. of Third Conference on Space Research*, Washington, May 1962, North Holland Co. (1963).

HEITLER, W., "The Quantum Theory of Radiation," Oxford University Press, 1944.

MARIANI, F., Manuscript in preparation (1963).

RISHBETH, H., LYON, A. J., and PEART, MARGARET, "Diffusion in the Equatorial F Layer," *J. Geophys. Res.*, **68**, 2559 (1963).

THOMAS, J. O., "The Electron Density Distribution in the F2 Layer of the Ionosphere in Winter," *J. Geophys. Res.*, **68**, 2707 (1963).

A MODEL OF THE QUIET IONOSPHERE

J. CARL SEDDON

Goddard Space Flight Center

Analysis of high-altitude rocket electron and ion density measurements suggests a simple model of the quiet ionosphere. Near $h_{max}F2$ the profile is given by the "α-Chapman" function, i.e., the Chapman electron density equation with $\sec\chi = 1$ and the "scale height" H a constant. A method is given for determining $h_{max}F2$, $N_{max}F2$, and H from $N(h)$ data obtained from ionograms. Well above the peak, the profile is taken to have a constant exponential slope of 200 km during the day and 150 km during the night. If simultaneous nearby measurements of the total electron content are available, a more accurate slope can be computed, thus providing the profile up to about 1000-km altitude.

INTRODUCTION

Electron density profiles above the F_2 maximum obtained with rockets and satellites were summarized by Wright (1960). He found that he could obtain an approximate agreement with these results if he used the Chapman electron density equation with H considered constant at 100 km and with sec $\chi = 1$. He computed the ratio of electron density above the maximum to that below, but these values tended to be a little higher than measured values. However, since the electron density results were mainly obtained from orbiting satellites and military rockets, the accuracy was not of the best. Subsequently, Berning (1960) obtained a sunrise profile to 1500 kilometers which when normalized did not agree well with the previous profiles that had been obtained. His results indicated the existence of a considerable gradient in the electron scale height. This profile was, however, due to unforeseen circumstances, obtained at ground sunrise when conditions could be changing at a rapid rate. A little later Bowles (1961) published a daytime profile obtained by the incoherent backscatter technique which indicated an approximate Chapman distribution with $H =$ about 90 km. His night profile indicates $H =$ about 75 km, although the densities close to the maximum are somewhat larger than a Chapman distribution with this scale height.

Nisbet and Bowhill (1960) published a series of profiles obtained with military rockets under difficult scientific conditions. They attempted to compare their normalized results with a Chapman distribution using a variable neutral scale height as given by Kallman (1959). The agreement was not particularly good. Berning (1961) revealed profiles showing a constant scale height H well above $h_{max}F_2$ of about 100 km in the daytime and 72 km during the evening.

Hanson and McKibbin (1961) reported ion density measurements that gave a value of H of 75 km in the evening for altitudes well above the maximum density. Pineo et al. (1961) reported on daytime results obtained by incoherent backscatter techniques which showed a scale height gradient existing similar to the sunrise results of Berning. More recently, Jackson and Bauer (1961) obtained a daytime profile with a value of H also given as 100 km.

This paper will discuss the electron and ion density results obtained with rockets under quiet ionospheric conditions and reasonably favorable scientific conditions. A simple model of the quiet ionosphere is obtained which can be expressed approximately in analytical form. This model is used to develop a means of obtaining from ionograms an approximate electron density profile and total electron content. It is also shown how such data used in conjunction with total electron

content measurements make possible the determination of the electron density profile above $h_{max}F_2$.

DISCUSSION

If the results of Berning (1961), Hanson and McKibbin (1961), and Jackson and Bauer (1961) are examined near the peak of the F_2 region, it is found that the measured electron densities do not follow a Chapman function with a scale height as high as 100 km, but these results are all in agreement that well above the peak altitude a constant exponential slope exists which, on the assumption that the ion and electron temperatures are equal, corresponds to a value of H of about 100 km in the daytime and about 75 km in the evening.

Yonezawa (1955a, b) showed that under the influence of vertical diffusion and nighttime attachment any $N(h)$ curve will at night tend to take the form of a Chapman layer of constant "scale height" H and sec $\chi=1$, referred to as an "α-Chapman" layer. Long (1962) showed that, if the $N(h)$ analysis of the ionogram includes an allowance for the underlying ionization that is usually neglected, then the profiles for all latitudes followed an α-Chapman variation to first order. This paper will also show that, for the limited daytime rocket data available, this also holds for quiet daytime conditions.

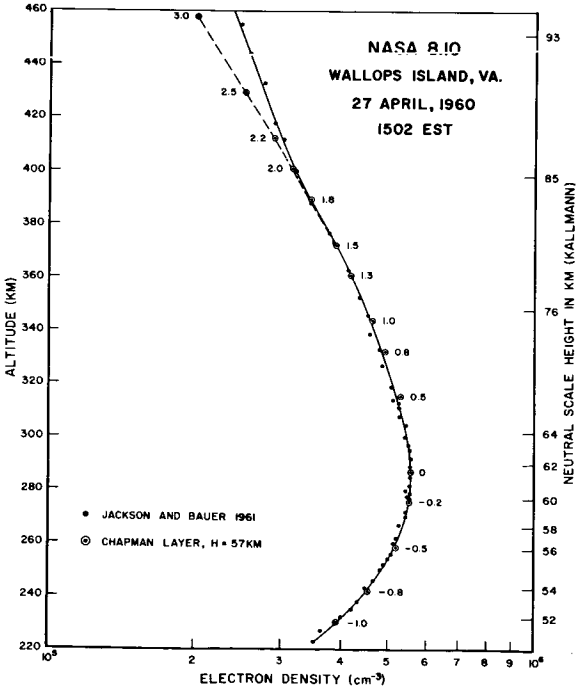


FIGURE 1.—Chapman distribution fitted to Jackson and Bauer electron density profile.

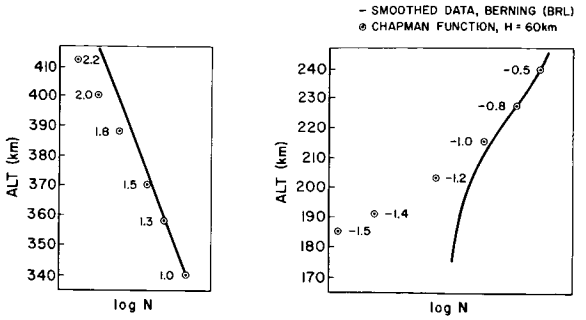


FIGURE 2.—Chapman distribution fitted to Berning electron density profile. Data from Wallops Island, Va., 0947 EST, July 13, 1960.

The results of Jackson and Bauer are plotted in figure 1 near the F_2 peak. The circled points show the α-Chapman function with a scale height of 57 km. The numbers represent the number of scale heights above or below the maximum. It is thus seen that there is indeed a very close agreement with this Chapman function over the range from -1 scale height up to about $+1.6$ scale heights with $h_{max}=287$ km. At 380 km, there is a break in the profile slope until a constant asymptotic slope is reached at an altitude of about 400 km. This is in disagreement with the result given by Jackson and Bauer that hydrostatic equilibrium begins at 350 km. Figure 2 shows a part of Berning's unpublished quiet-day results which also indicates that equilibrium begins at 400 km. His results follow a Chapman distribution very accurately from -1 to $+1.5$ scale heights with $H=60$ km and $h_{max}=280$ km. Figure 3 shows the nighttime ion density results of Hanson and McKibbin which also show equilibrium beginning at 400 km. Their results follow a Chapman distribution from below -1.2 to $+1.3$ scale heights with $H=43$ km and $h_{max}=309.5$ km. Berning's unpublished night flight at Wallops Island, Virginia, at 2143 EST July 13, 1960, also follows the Chapman function very accurately from -2.0 to $+1.8$ scale heights with $H=63$ km and $h_{max}=357$ km.

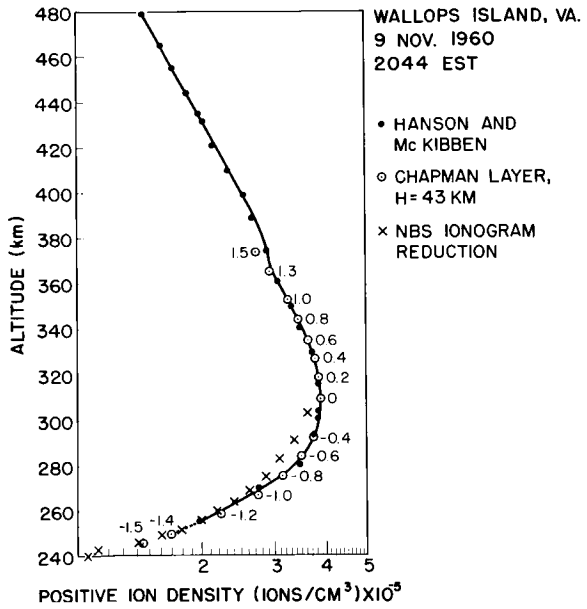


FIGURE 3.—Chapman distribution fitted to Hanson and McKibbin ion density profile.

It is difficult to state precisely at what altitude the neutral particle scale height has a value equal to H, but it would seem reasonable to suppose that it is near to $h_{max}F_2$. Kallmann's (1961) average daytime scale heights are shown on the right-hand side of figure 2. The agreement is even better when compared with Berning's daytime flight. Her nighttime average value agrees reasonably well with Berning's summer night firing, but the agreement is rather poor when compared to Hanson and McKibbin's November night firing.

These data thus indicate that under quiet conditions the F_2 distribution is like α-Chapman distribution from -2.0 to $+1.5$ scale heights at night and from -1.0 to $+1.5$ in the daytime. Above 1.5 scale heights there is a transition region of about 25 kilometers after which diffusive equilibrium exists, with a constant exponential slope of about 150 km in the evening and 200 km in the daytime. Data late at night are not available, and it is likely that before sunrise the slope may decrease to lower values. A method will be suggested to check on this.

Model of the Quiet Ionosphere

As the ionosphere decreases the tracking accuracy of systems using radio waves, it would be desirable to be able to express the profile in terms of approximate analytical functions to simplify calculations. Therefore, the ionization in the E region and above was reviewed, and it was found that a reasonable approximation in the daytime below -1.0 scale height may be had by drawing a straight line on semilog paper to $N_{max}(E)$ at an altitude of 100 km. Though the actual profiles show variations around this assumed profile, the electron contents are about equal. Figure 4 shows the approximate model for the quiet middle-latitude ionosphere. Table 1 shows an approximate representation of the profiles in analytical form which may be useful in refraction calculations. S is obtained from a knowledge of the initial and final values. Also shown are formulas for the electron content. It is assumed here that the nighttime region may be taken to be entirely Chapman-like below h_{max}. If appreciable sporadic E exists, it may be assumed to have a plasma frequency equal to the lowest F-reflection frequency plus one-half the gyrofrequency and an average thickness of 1 km (Seddon, 1961).

Table 2 shows the electron content calculated from the formulas of table 1 compared to the values obtained from numerical integration of rocket data with the assistance in some cases of E-region data obtained from $P'-f$ analysis performed by the National Bureau of Standards.

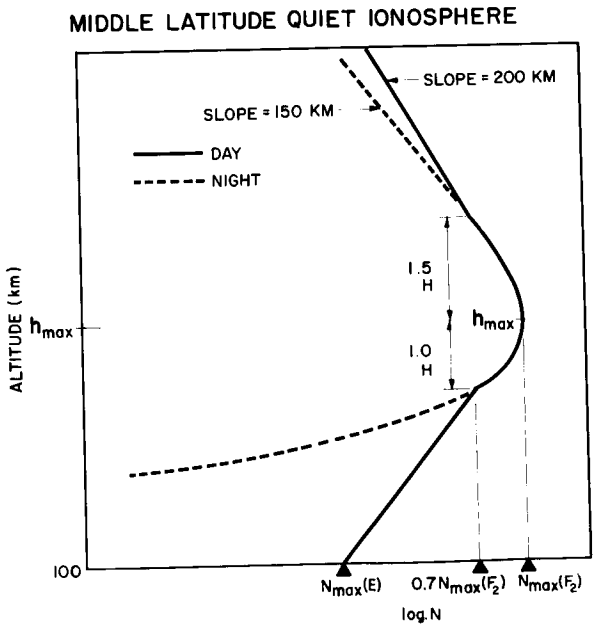


FIGURE 4.—Simple model of the quiet ionosphere.

TABLE 1.—*Middle-Latitude Quiet Ionosphere*

Night N	$N_{max}(F_2)\mathrm{Ch}(Z)$	$Z \leq 1.5$, $Z=(h-h_{max})/H$
	$0.7N_{max}(F_2)\exp-[(h-h_{max}-1.5H)/150]$	$h \geq h_{max}+1.5H$, $h=$ altitude
Day N	$N_{max}(E)\exp(h-100)/S$	$N_{max}(E) \leq N \leq 0.7N_{max}(F_2)$, $100 \leq h \leq h_{max}-H$
	$N_{max}(F_2)\mathrm{Ch}(Z)$	$-1 \leq Z \leq 1.5$
	$0.7N_{max}(F_2)\exp[-(h-h_{max}-1.5H)/200]$	$h \geq h_{max}+1.5H$
Night N_b	$1.312HN_{max}(F_2)$	
Night N_a	$1.30HN_{max}(F_2)+150\times 0.7N_{max}(F_2)$	
Night N_t	$2.61HN_{max}(F_2)+150\times 0.7N_{max}(F_2)$	
Day N_b	$S[0.7N_{max}(F_2)-N_{max}(E)]+0.88HN_{max}(F_2)$	
Day N_a	$1.30HN_{max}(F_2)+200\times 0.7N_{max}(F_2)$	
Day N_t	$S[0.7N_{max}(F_2)-N_{max}(E)]+2.18HN_{max}(F_2)+200\times 0.7N_{max}(F_2)$	

On comparing the rocket results with the results obtained by the National Bureau of Standards using ionograms from a nearby ionosonde, it was found that, though agreement was generally good, at times considerable disagreement existed regarding H and/or h_{max}. The National Bureau of Standards technique involves a parabolic assumption near h_{max} which should give fairly good results if the ionogram can be read to frequencies very close to f_0F_2. Below about one-half scale height, however, the parabolic approximation becomes rapidly poorer, and examination of their data frequently shows few of no points higher than $-\frac{1}{2}$ scale height. In addition, the method depends on an accurate value of f_0F_2. The same $N(h)$ data were used to obtain the best fit possible to a Chapman distribution, and results were obtained giving in general better agreement with the rocket results, including $N_{max}F_2$ which is obtained without the use of f_0F_2.

Experimental Method

Though various methods can be devised to fit the $N(h)$ values obtained, the following method, which can be performed on a desk computer or an electronic computer, was used in this paper. Both were actually used, the electronic computer being the relatively slow LG-30 digital computer.

All of the values obtained from the $N(h)$ reduction were used that were within one scale height below the estimated h_{max} in the daytime and within about 1.5 scale heights below h_{max} at night. The highest altitude was called h_1, the next h_2, etc. If the points fall on a Chapman curve, the relation

$$H_1=(h_1-h_i)/(Z_1-Z_i) \tag{1}$$

must be constant for each value of i, where Z is the solution of $\mathrm{Ch}(Z)$ with sec $\chi=1$, i.e.,

$$N=N_{max}\exp\tfrac{1}{2}(1-Z-e^{-z}) \tag{2}$$

The values of Z can be found from the relation

$$\mathrm{Ch}(Z_i)=N_i/N_{max} \tag{3}$$

The computer is given a value of N_{max}, a little less than the value computed from the observed f_0F_2. The h values are obtained from the $N(h)$ reduction, along with the corresponding value of N. The result obtained is a series of values H_{12}, H_{13}, etc., which will increase steadily in value if the N_{max} used is too small and if the $N(h)$ values are, in fact, truly Chapman. Due to the fact that there is some variation from a true Chapman function, the variation will not be entirely smooth. The mean value $\bar{H}$ is computed and also the average deviation from the mean.

A slightly large value of N_{max} is then given to the computer, and the same computations are made again. If the $N(h)$ points were exactly on a Chapman curve, a value of N_{max} would ultimately

TABLE 2.—*Summary of Results*

Date, time, place*	Data source	H, km		h_{max}, km		N_{max}, $10^5 cm^{-3}$		N_b, $10^{12} cm^{-2}$	Calculated N_b†	N_a, 10^{12} cm^{-2}	Calculated N_a†	N_a/N_b	Calculated N_a/N_b
		Parabolic	Chapman	Observed	Chapman	Observed	Chapman						
1	Rocket	------	57.0	287.0	287.0	5.60	5.60	5.87	6.00	11.85	12.00	2.02	2.0
2	$P'-f$	59.3	60.3	299.0	294.0	5.64	5.57	6.51, 6.24‡	6.15	------	12.16	------	2.0
3	$P'-f$	49.7	59.0	281.0	284.3	5.10	5.13	5.37, 5.55‡	5.60	------	11.11	------	2.0
4	Rocket	------	63.0	357.0	357.0	4.60	4.60	3.91	3.80	8.60	8.59	2.20	2.26
5	$P'-f$	49.6	63.0	363.0	370.7	4.54	4.605	3.27, 3.60‡	3.81	------	8.60	------	2.26
6	Rocket	------	43.0	310.0	309.5	3.90	3.90	2.14	2.20	6.89	6.27	3.22	2.9
7	$P'-f$	43.9	50.5	318.0	321.7	3.72	3.76	2.34, 2.48‡	2.49	------	6.41	------	2.6
8	Rocket	------	60.0	280.0	280.0	5.03	5.03	5.77	5.50	10.96	10.96	1.90	2.0

*April 27, 1961: 1—1502 EST, Wallops Island.
2—1500 EST, Wallops Island.
3—1500 EST, Fort Belvoir.
July 13, 1960: 4—2143 EST, Wallops Island.
5—2145 EST, Fort Belvoir.
Nov. 9, 1960: 6—2044 EST, Wallops Island.
7—2044 EST, Wallops Island.
July 13, 1960: 8—0945 EST, Wallops Island.

†Calculated from table 1.

‡Chapman function above $N(h)$ reduced data.

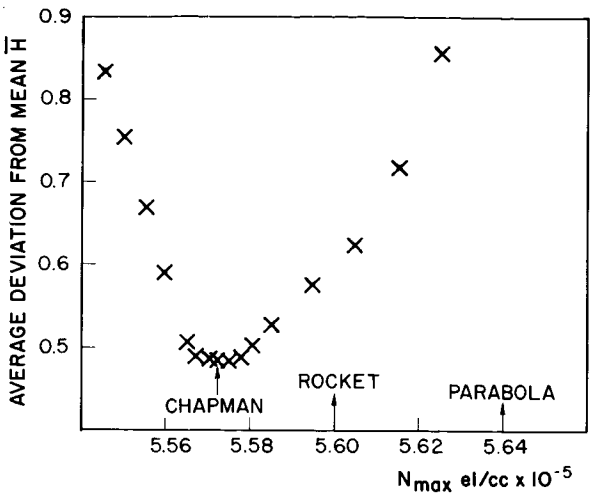


FIGURE 5.—Determination of $N_{max}(F_2)$ by Chapman-fit to ionogram $N(h)$ reduction. Data from Wallops Island, Va., 1500 EST, April 27, 1961.

be found where the average deviation would be zero. Due to irregular variations from true Chapman, the average deviation decreases rapidly and nearly linearly to a value dependent on the size of such variations and then increases again as illustrated by figure 5. $h_{max}F_2$ is the average value of the values obtained for $h_{max\ i}$ for the correct N_{max}, where

$$h_{max\ i}=h_i-\bar{H}Z_i \qquad (4)$$

DISCUSSION OF RESULTS

The values obtained for H, h_{max}, and N_{max} are shown in table 2 in the columns labeled "Chapman." The values obtained by NBS using the parabolic approximation are shown in the other columns. The table is arranged to facilitate comparison with the rocket results. The Wallops Island ionosonde is only a few miles from the rocket launcher. The Fort Belvoir ionosonde is approximately 160 km to the west, or about 8 minutes earlier in solar time.

The values for either or both H and N_{max} obtained by the Chapman technique in general agree better with the rocket results than those obtained from the parabolic approximation except for the one case of the winter night firing. It is quite likely that the ion trap measurements are not as accurate as the electron density measurements, since the $N(h)$ reduction disagrees by a variable amount up to nearly 10%, which can introduce a considerable change in the value of H.

In all the other cases examined, it was found that the $N(h)$ reductions were in good agreement with the rocket data except that, less than one scale height below the maximum, the values of N were always slightly lower. However, this slight decrease tends to make both H and $h_{max}F_2$ too large and $N_{max}(F_2)$ very slightly too small. The result is to cause N_b to be too large in the daytime. A better result is obtained by integrating the $N(h)$ results up to -1 scale height and then using the Chapman profile to peak. At night, the ionization at the lower altitudes not determined by the ionogram reduction technique more than offsets this effect. It is more accurate at night (and easier) to assume that the region is Chapman-like rather than to integrate the profile obtained by the reduction.

Figure 5 shows some variation in $N_{max}F_2$ obtained by the different measurements. At the time of the rocket firing, the value of $N_{max}F_2$ at Fort Belvoir by the Chapman-fit method was 5.13×10^5 el/cc, whereas the Wallops Island ionosonde gave 5.57×10^5 el/cc. This indicates a horizontal gradient in the electron density, increasing to the east. The rocket firing was approximately east of the ionosonde and the measured value was 5.60×10^5 el/cc. If the observed f_0F_2 at Wallops Island is assumed to be correct, then the horizontal gradient would have to be in the opposite direction east of the ionosonde.

Another interesting observation is that the electron density profiles, both day and night, show a departure from the Chapman profile close to 225 km. Also, if the data from the daytime ionograms is used to calculate $h_{max}F_2-H$, or from the nighttime ionograms $h_{max}-2H$, a value around 230 km is usually obtained. It would thus seem that the strong tendency to assume a Chapman distribution exists above about 225 km but not below it.

The ionogram data for July 13, 1960, was not included in table 2. Although the day was not classified as stormy, the F_2 region was undergoing rather rapid changes with time. None of the Fort Belvoir ionograms near the time of the rocket firing agreed with the rocket data. The values obtained for H varied at times at a rate of 1 km per minute. Sometimes only 3 points were avail-

able within one scale height of the peak. Under such circumstances we cannot conclude that the profile was truly like a Chapman profile. It was noted, however, that the values of f_0F_2 computed on the assumption that it was like a Chapman profile gave values that were from 1% to 2% larger than the observed f_0F_2. It was noted that for these instances the recorded virtual heights were abnormally large, sometimes exceeding 600 km. This was presumably due to the fact that f_0F_2 was only slightly larger than f_0F_1. This results in considerable absorption and a rapid rate of change in virtual height with frequency near f_0F_2. This raises a question as to whether it is possible to measure f_0F_2 accurately from such an ionogram. Additional evidence is seen in the fact that during the period 20 minutes before firing time to 80 minutes after, during which 10 ionograms were taken, only one ionogram (20 minutes after firing) showed f_0F_2 at Fort Belvoir to be greater than the rocket measured value at Wallops Island, with only two (30 and 40 minutes after firing) being nearly as large.

If the ionogram reduction shows a Chapman distribution, one can obtain a reasonably accurate value for N_b, the electron content below the maximum. Then if measurements are made nearby of the total electron content by means of satellites or two-frequency moon echoes, it is a simple matter to compute the value of the constant exponential slope above $h_{max}F_2$. The rocket values of N_a+N_b were assumed to be known, and the calculation of these slopes agreed quite well in every case. It thus seems to be feasible to make measurements of the slope by this technique under quiet conditions if this is not done near sunrise or sunset.

The ratios N_a/N_b are quite interesting. Spring afternoon, summer morning, and summer evening values are all fairly close to 2.0. Garriott (1960) also finds a similar value for autumn and winter days. On autumn and winter nights, he finds the ratio to be as high as 4. His two winter evening measurements made at 2044 local time are in very good agreement with the value of 3.2 reported here for that time on November 9. The model suggested in this paper shows that the ratio will increase if the value of H near h_{max} decreases to lower values with no significant change occurring in the constant logarithmice slope above h_{max}. As the ionosphere closely approximates a Chapman region on the lower side at night, and as the ratio for a true Chapman region is 2.14 (Wright, 1960), we would expect that the ratio would be greater than this whenever the value of H is less than one-half the exponential slope; i.e., about 75 km in the evening.

Regarding disturbed ionospheric conditions, only a few remarks can be made. Berning (1961) reported on a flight to 400 km at Wallops Island under disturbed conditions. f_0F_2 was below normal and $h_{max}F_2$ was higher than normal. The F_2, region was very close to a Chapman distribution from $-\frac{1}{2}$ scale height to the peak of the flight, $+1.3$ scale heights. f_0F_1 was nearly equal to f_0F_2, and under such conditions it is difficult to obtain an ionogram reduction close to the peak. Also, the shape of the profile above $+1.3$ scale heights is not yet known under such conditions.

CONCLUSIONS

The $N(h)$ reduction of ionograms can be easily checked to determine whether the F_2 region follows an α-Chapman distribution. If it does, the scale height H, the altitude $h_{max}F_2$, and electron content N_b can be determined with more consistent accuracy by using a best fit to a Chapman function than by using the parabolic technique. In addition, it is possible to predict accurately the electron density profile to $1.5H$ above $h_{max}F_2$ and approximately the profile above this altitude and therefore obtain a fair value for the electron content N_a during the daytime and evening. If the total content is measured by some method, the electron distribution can be predicted to nearly 1000 km using ionogram data and the suggested model with the required logarithmic slope obtained from the total content measurement.

ACKNOWLEDGMENTS

I am greatly indebted to Dr. W. W. Berning of the Ballistic Research Laboratories for making available the detailed results of his rocket firings before publication. Thanks are also due Dr. W. B. Hanson and Dr. D. D. McKibbin of the Lockheed Aircraft Corporation and Mr. J. E. Jackson and Dr. S. J. Bauer of the Goddard Space Flight

Center for tabulated values of their results. $N(h)$ reductions were kindly supplied by Mr. J. W. Wright of the National Bureau of Standards who also contributed many helpful discussions. The computations were performed by Mrs. V. E. Fankhauser of the Goddard Space Flight Center.

REFERENCES

BERNING, W. W., *J. Geophys. Res.*, **65**(9), 2589, 1960.
———, paper presented at URSI meeting, May 1961.
BOWLES, K. L., *J. Res. NBS*, **65**D(1), 1, 1961.
GARRIOTT, O. K., *J. Geophys. Res.*, **65**(4), 1151, 1960.
HANSON, W. B., and MCKIBBIN, D. D., *J. Geophys. Res.*, **66**(6), 1667, 1961.
JACKSON, J. E., and BAUER, S. J., *J. Geophys. Res.*, **66**(9), 3055, 1961.
KALLMANN, H. K., *J. Geophys. Res.*, **64**(6), 615, 1959.
———, *J. Geophys. Res.*, **66**(3), 791, 1961.
LONG, A. R., *J. Geophys. Res.*, **67**(3), 989, 1962.
NISBET, J. S., and BOWHILL, S. A., *J. Geophys. Res.*, **65**(11), 3609, 1960.
PINEO, V. C., et al., paper presented at URSI meeting, Washington, D.C., May 1961.
SEDDON, J. C., NASA Tech. Note D-1043: Sporadic-E Monograph, edited by Matsushita and Smith, Pergamon Press, New York (to be published).
WRIGHT, J. W., *J. Geophys. Res.*, **65**(1), 185, 1960.
YONEZAWA, T., *J. Radio Res. Lab. Japan*, **2**, 125, 1955*a*.
———, *J. Radio Res. Lab. Japan*, **2**, 281, 1955*b*.

THE UTILIZATION OF IONOSPHERE BEACON SATELLITES

G. W. SWENSON, JR.

University of Illinois

This paper discusses the application of beacon transmitters in earth satellites to ionospheric research, and the need for worldwide cooperation in beacon satellite observation. The theories of Faraday rotation and Doppler methods of determining atmospheric electron content are reviewed briefly and the working formulas are listed. Desirable characteristics of beacon satellites are outlined. Methods of observation and equipment are discussed in some detail, as are the minimum standards for the production of high-quality data. An extensive bibliography is also presented.

INTRODUCTION

With the launching of the first artificial satellite Sputnik I (1947 α), an opportunity was presented to the scientific community to investigate the ionosphere by observing the changes in character of a radio signal, produced by its passage through this region. The early Russian satellites transmitted signals at frequencies of approximately 20 and 40 Mc, low enough to be influenced substantially by existing electron densities and magnetic fields in the ionosphere. Sputnik III (1958 $\delta 2$) was particularly useful for ionospheric studies because its rather eccentric orbit carried it, at times, well above the most heavily ionized regions, and because it transmitted for well over a year.

Three U. S. satellites have transmitted radio signals in the vicinity of 20 Mc: Explorer VII (1957 $\iota 1$), and Discoverers XXXII (1961 $\alpha\gamma 1$) and XXXVI (1961 $\alpha\kappa 1$). The former, unfortunately, had an elliptically polarized antenna and was spin stabilized. Therefore it was of limited usefulness for ionospheric research. The latter two were specifically designed to study scintillation, had short lifetimes, and orbited at rather low heights. Therefore they too had limited utility.

With the exception of the two Discoverers mentioned above (the Nora Alice experiment), no satellites specifically designed for ionosphere research by means of their radio transmissions have been available. Two efforts by the National Aeronautics and Space Administration to launch "ionosphere beacon" satellites were unsuccessful.

A new NASA program is being implemented which aims at providing such a satellite with characteristics optimized for investigations of the ionosphere-beacon variety. It is the purpose of this paper to encourage wide utilization of this beacon satellite for ionospheric research and to suggest some ways in which this can be done.

At present a number of satellite tracking stations are actively engaged in ionospheric research by means of satellite beacon transmitters. The stations known to have been active at some time since 1957 are shown on the map in figure 1. This map is undoubtedly incomplete, but it is quite apparent that the geographic coverage is grossly nonuniform. Of the stations charted, few have reported the results of long series of observations and many apparently have been operated only on an ad hoc basis.

This method of ionospheric research is relatively simple and inexpensive, and it seems highly desirable that a more widespread geographic coverage be obtained and that observations be made by more-or-less standard techniques over long periods of time.

This report is intended as a type of handbook for persons contemplating field programs. It is hoped that it will help promote participation on a widespread geographical basis. To this end, atten-

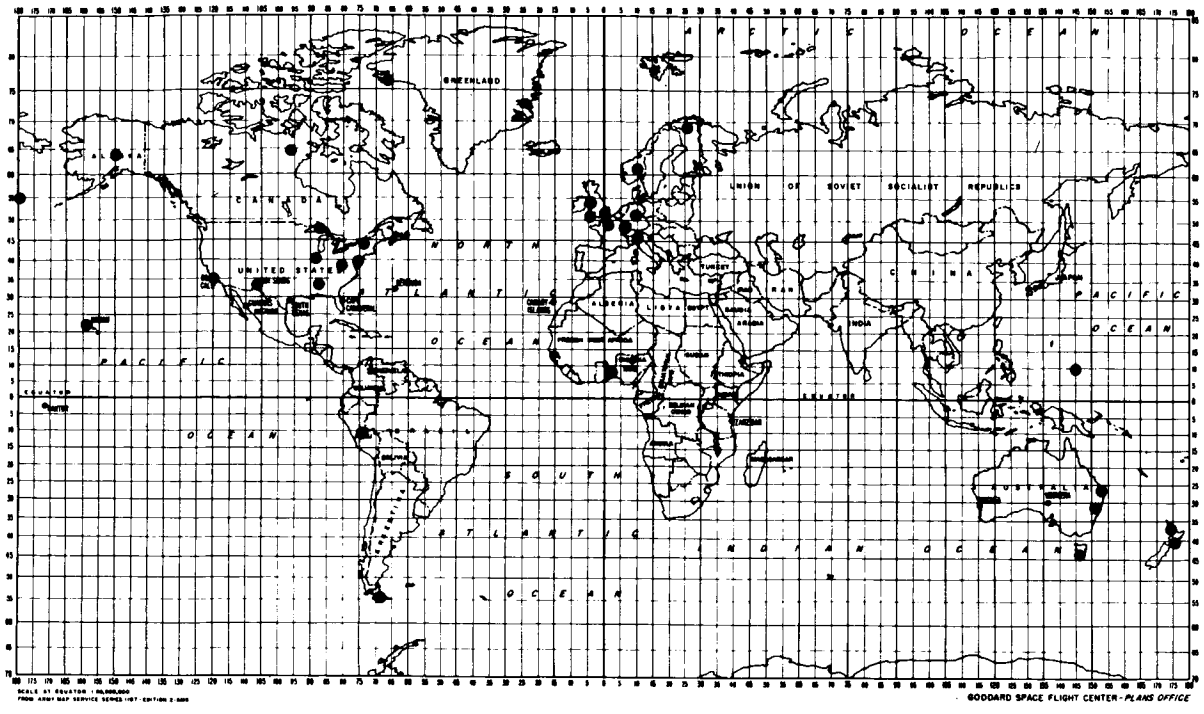


FIGURE 1.—Worldwide distribution of known ionosphere-beacon satellite monitoring stations.

tion will be concentrated upon the simplest feasible methods and equipment, with particular emphasis upon operation at remote sites. The basic requirements are an antenna, a receiver, a recorder, and a timing system. These requirements are essentially the same whether one considers a station operated at a laboratory with elaborate equipment and highly trained personnel or at a remote field site having only absolutely necessary facilities. Principal requirements for this sort of operation are a high degree of reliability, adequate technical performance, and the nearest practicable approach to automatic operation. Only if these criteria are met can records of good quality be obtained over long periods by relatively inexperienced personnel. It is also important to a synoptic program that data taken at different locations and times have comparable characteristics. Therefore, standards for data collection will be discussed in a later section of this report.

SCIENTIFIC OBJECTIVES OF THE BEACON TRANSMITTER PROGRAM

An orbiting beacon provides the means for a synoptic study of the ionosphere as it varies in time and space. One of the principal features of this research technique is the simplicity of the equipment and of the methods of data reduction. Another is the ease with which observations can be made over a very long period.

The collected data can be used to determine the columnar electron content (or integrated electron density) and the distribution of irregularities in the ionosphere. Although the columnar electron content is not generally regarded as a parameter directly applicable to engineering problems, it is useful in interpreting other data, such as ionograms; it serves as a check on determinations of the electron density profile; and it is readily correlated with other geophysical phenomena such as magnetic activity. The general method of determining columnar electron content has been demonstrated by several investigators and may be regarded as well established. It is therefore appropriate to undertake a long term program of determining temporal changes in the columnar electron content.

A beacon satellite provides a unique means for studying small scale irregularities in the ionosphere. Heretofore, radio stars have provided such a means, but they are of limited utility in

determining the latitude distribution of scintillation and do not possess the advantages of single-frequency phase-coherent signals. At the time of this writing, the state of understanding of even the "morphology" of scintillation is sketchy and unsatisfactory, and virtually no progress has been made toward the reconciliation of theory with observations.

Several authors have published the results from past determinations of the integrated electron content obtained by the use of satellite signals.[1–13] Thus, the techniques of observation and analysis have been tested, and a number of them have been found suitable for operational use. One of the most striking features of the published data is the range of variation in the electron content. Typical diurnal variations over a range of four or five to one occur between midday and midnight. A marked seasonal variation is also apparent, and geomagnetic disturbances appear to be accompanied by very substantial changes in electron content.

These effects should be studied in more detail. Sufficient data should be collected to permit the comparison of results during different parts of the sunspot cycle. Many more data are needed, from a wide geographical distribution of receiving stations, to give a better understanding of the latitude and longitude dependence of electron content.

The distortion of radio signals from satellites by random irregularities has been widely studied, but the result has been only a rather meager understanding of the phenomenon.[14–27] It is known that scintillation varies strongly with local time and with latitude (probably being associated with geomagnetic latitude) and that it is somehow associated with other geophysical phenomena such as magnetic activity and spread F. In north temperate regions, scintillation caused by irregularities in the F region apparently occurs mostly at night and north of geomagnetic latitude 50°N. This latitude distribution is still uncertain and needs further study, particularly between about 120°W. and 50°E. In the Southern Hemisphere, the corresponding northern boundary has been observed only recently and needs further study. If, as seems logical, F region scintillation is associated with the auroral zones, it may be expected that activity is somewhat less in the vicinities of the magnetic poles. Observations of high frequency signals from polar regions are as yet too few to permit resolution of this question.

Some efforts have been made to determine the altitudes at which irregularities responsible for scintillation occur. While results to date suggest heights in the vicinity of 300 km, it is not known how these heights vary with time or geographic position.

Scintillation is sometimes noted on the equatorial side of the temperate zone boundaries, usually in the daytime. This phenomenon has been little studied and is not well understood. Also, it is known that scintillation is prevalent in equatorial regions, but there have been no reports of systematic studies.

Scintillation appears to be much less prevalent and of much lower intensity at present than it was in 1958–59, but insufficient data are now being collected to permit accurate evaluation of the relationship between scintillation and the sunspot cycle. Clearly, it is highly desirable that a long-term effort be instituted to gather scintillation data, involving many observing stations with a wide geographical distribution.

Irregularities in electron density of the order of hundreds of kilometers in extent have been reported by various researchers.[9,11] Satellites in low orbits are very useful for studying such phenomena, since they cover large regions in short periods of time.

Many reports of unusual radio propagation phenomena observed by means of satellites have been published.[18,28–35] Most of these observations were from low satellites which were within the F region; however, it is of scientific interest to determine whether the antipodal propagation effect, for example, also occurs for satellites in higher orbits.

DESIRABLE CHARACTERISTICS OF A BEACON TRANSMITTER

To give the widest geographical coverage and permit studies of the polar ionospheres, the satellite orbit should be inclined approximately 90° from the earth's equatorial plane. In order to minimize certain difficulties in data analysis and

to keep the satellite well above the regions producing scintillation and above most of the electrons in the ionosphere, the height should be at least 1000 km. A circular orbit also simplifies analysis. An orbit meeting these criteria will be relatively free from atmosphere drag, permitting accurate prediction of observation times and simplifying the computation of accurate post-factum ephemerides.

In the past considerable difficulty has been experienced as a result of spinning or tumbling of satellites. It is desirable either that the antennas have electromagnetic symmetry about the spin axis or that the satellite be stabilized in such a way that the antenna aspect presented to a ground observer changes only very slowly, say at one revolution in several minutes, or less.

For reasons that will be apparent later, it is very useful to have several transmitting frequencies and it is important that they be locked together in phase. Obviously, enough power should be radiated to permit reception with good signal-to-noise ratios by simple equipment on the ground. For the orbits under consideration, a few hundred milliwatts is sufficient for high frequency radiation whereas 50 mw or more is needed in the VHF band.

The emissions should be purely continuous-wave in nature, with no modulation which would confuse a phase-locked receiver. A sequence of frequencies which has been chosen as appropriate for such an experiment is 20.005, 40.010, 41.010, and 360.080 Mc, all harmonics of 1.00025 Mc. It is expected that future ionosphere beacon satellites of the United States will utilize these frequencies.

Beacon experiments have been performed often enough so that in the future they should be considered "monitor" rather than "definitive" experiments. Thus, the satellite should transmit for the longest possible period. If the desired orbital characteristics are attained, the satellite will have an orbital lifetime of many years; and through use of solar power supplies it is also feasible to plan for a comparable transmitter lifetime. To guard against the nuisance of a transmitting satellite whose usefulness has ceased, a reliable means of turning off the transmitter must be provided; but it should also be possible to turn it on again if the occasion demands. It is desirable that the actual experiment be conducted during at least half of a sunspot cycle, and preferably longer.

THE FARADAY ROTATION METHOD OF MEASURING INTEGRATED ELECTRON DENSITY

Basis of the Measurements

In traversing a magnetoionic medium a plane wave undergoes a rotation of its plane of polarization, the total angle of rotation depending approximately on the average magnetic field component in the direction of propagation and on the total number of electrons in a column one square meter in cross section and equal in length to the length of the propagation path. As a satellite moves continuously with respect to the magnetic field, there occurs a time-variation of the total angle of rotation that the electric vector experiences along the propagation path. The result is a continuous rotation of the electric vector at the receiving point. The magnetic field of the earth is known to the desired accuracy; therefore, the total angle of rotation along the transmission path determines, approximately, the columnar electron content along the path. In practice it generally is not possible to measure the total angle of rotation; instead, the *rate* of rotation and the number of complete rotations between two points on the orbit are determined. Methods have been developed to determine the total electron content from these data.

The principal advantage of the Faraday rotation method of determining electron content is its simplicity, both in data collection and in analysis. Other methods may yield more detailed and accurate data, but at the cost of far greater effort and expense. The Faraday rotation method seems well suited for long-term monitoring experiments.

"Single Frequency" Methods

Measurement of the rate of change of polarization angle at a given instant yields an approximate determination of the electron content. Theoretically, a determination can be made for a given instant of time (that is, for a given ray path). In practice it is necessary to measure the number of revolutions of the electric vector in a

finite time in order to determine the rate accurately enough. This *rate of rotation method* probably constitutes the simplest means of obtaining an approximation of the electron content of the ionosphere.

The basic equation, as derived by Bowhill,[36] is

$$\int N\,dh = \frac{zf^2}{K_1 H_x v_x}\,\dot{\Omega}\,, \qquad (1)$$

where

$\int N\,dh$ = number of electrons in a column 1 meter square extending from the satellite to the receiver;

$$K_1 = \frac{e^3 \mu_0}{8\pi^2 m^2 C \epsilon_0} = 0.0297 \text{ (mks rationalized units)};$$

z = altitude of the satellite (m);
f = radio frequency (cps);
H_x = component of the earth's magnetic field in the direction of v_x (amp/m);
v_x = horizontal component of the satellite velocity with respect to the observer m/sec);
$\dot{\Omega}$ = rate of change of the Faraday rotation angle (rad/sec).

This formula was derived under the assumptions of a horizontally moving satellite, a plane earth, and a horizontally stratified ionosphere. Garriott[7] has discussed the errors involved in neglecting the vertical component of velocity and has developed a method of correction. In addition, it is assumed that the frequency is high compared with the maximum critical frequency of the ionosphere, and that the ray path is a straight line. All these assumptions lead us to expect limited accuracy from the method, and most observers have preferred to resort to more accurate formulas requiring more elaborate observational techniques. However, the regular and sporadic variations in electron densities observed with the more accurate methods are large compared with the errors expected to arise from the aforementioned assumptions. In fact, detailed comparison of data (8) reduced by equation 1 with data reduced by more accurate methods[12] shows that the major features are the same. Thus equation 1, whose application does not require the use of an electronic computer, is useful for the quick evaluation of large amounts of data.

The total angle of Faraday rotation is given approximately[37] by

$$\Omega \approx \frac{K_1}{f^2}\,\bar{M} \int N\,dh\,, \qquad (2)$$

where

Ω is the total angle of rotation of the electric vector in traversing the ionosphere,
H is the magnetic field (amp/m),
ϕ is the angle between H and the ray,
i is the vertical angle of the ray, and
$\bar{M}$ is a suitable average value of $H \cos\phi \sec i$.

Unfortunately, under typical conditions Ω amounts to many complete revolutions so that the integral in equation 2 can be determined only ambiguously. However, if the ionosphere is assumed to have no horizontal gradients and if the satellite is above most of the electrons in the ionosphere, then $\int N\,dh$ will be the same for any path. Consider the satellite to move from point 1 to point 2. The difference $\Omega_1 - \Omega_2$ is the differential rotation of the electric vector seen by the observer during this interval. The integrated electron density can then be determined unambiguously as

$$\int N\,dh = \frac{\Omega_1 - \Omega_2}{\bar{M}_1 - \bar{M}_2}\,\frac{f^2}{K_1}\,, \qquad (3)$$

in what is therefore called the *differential rotation method*. Here $\bar{M}$ must be computed carefully from measured values at ground level. A computer program has been prepared for this purpose[12,38] which can be used to compute the necessary values of $\bar{M}$ for any satellite position as viewed from any geographical location.

Second-Order Faraday Rotation Methods

The differential rotation method involves essentially the same assumptions and approximations as the rate of rotation method. It is hoped that future ionosphere beacon satellites will be in high, circular orbits so that the effects of neglecting the vertical component of velocity are minimized. Corrections can be made for the high-frequency approximation and for refraction by using two widely separated frequencies and re-

ducing the data by means of second-order formulas given by Yeh.[39] It is convenient (but not necessary) for this purpose to use two harmonically related frequencies, say 20 and 40 Mc. The use of these methods indicates that the errors caused by refraction and the high-frequency approximation are in the neighborhood of 5% for close passages and up to 30% for distant passages.

In the *second-order rate of rotation method* the integrated electron density is given[39] by

$$\int N\,dh = \frac{zf^2}{K_1 H_x v_x} \frac{1}{m^2-1}\left[m^4\dot{\Omega}(mf) - \dot{\Omega}(f)\right], \quad (4)$$

in which m is the ratio of the higher to the lower frequency and the other symbols are as defined before, $\dot{\Omega}(mf)$ and $\dot{\Omega}(f)$ being the rates of Faraday rotation of the two frequencies.

A similar second-order extension may be made of the *differential rotation method.* The integrated electron density is given by

$$\int N\,dh = \frac{f^2}{K_1(m^2-1)} \frac{m^4\Delta\Omega(mf) - \Delta\Omega(f)}{\bar{M}_1 - \bar{M}_2}, \quad (5)$$

in which $\Delta\Omega(mf)$ and $\Delta\Omega(f)$ are the rotations (in radians) of the frequencies mf and f observed as the satellite moves from point 1 to point 2, and $\bar{M}_1$ and $\bar{M}_2$ are the magnetic field geometric factors for the two points.

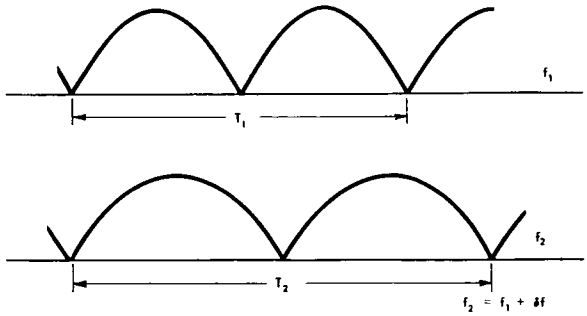


FIGURE 2.—Typical received signal obtained in the "Closely Spaced Frequencies" method, employing plane-polarized antennas and receivers with logarithmic amplitude response.

"Closely Spaced Frequencies" Method

The use of two closely spaced frequencies permits an unambiguous determination of the total angle Ω through which the electric vector is rotated in traversing the ionosphere. By using plane-polarized antennas and receivers with logarithmic amplitude response, records such as those shown in figure 2 will be obtained. It is most convenient to reckon in terms of complete revolutions, as indicated by the nulls in the pattern; a null occurs every half revolution. The working formulas (referred to figure 2) are:

$$\delta\Omega = \frac{2\pi(T_2 - T_1)}{T_1} + n\pi \text{ radians}, \quad n = 0, 1, 2, \ldots ; \quad (6)$$

$$\int N\,dh = \frac{f^2\,\Omega}{K_1 \bar{M}} ; \quad (7)$$

$$\frac{\partial\Omega}{\partial f} = \frac{-2\Omega}{f} \approx \frac{\delta\Omega}{\delta f} ; \quad (8)$$

$$\Omega = -\frac{f}{2}\frac{\delta\Omega}{\delta f} . \quad (9)$$

The electron content is determined by substituting from equations 6 and 9 into equation 7. This is a first-order determination; however, it is anticipated that it will be used at about 40 Mc, at which frequency the refraction and high-frequency approximation errors should be quite small.

The principal advantages of this method are simplicity in the data reduction, and the ability to make an unambiguous measurement of integrated electron content in a relatively short interval of time. The latter feature is important in evaluating horizontal variations of electron distribution.

THE DOPPLER METHOD OF MEASURING INTEGRATED ELECTRON DENSITY

The high velocity of a satellite relative to a ground-based observer results in a substantial Doppler shift in the observed frequency. In the ionosphere, the phase velocity depends upon the electron density; thus, the observed frequency also depends upon this parameter. Observation of the exact frequency as a function of time can be used to determine the electron content of the ionosphere.

Many authors have discussed the Doppler method of electron content measurement and several series of actual measurements have been reported.[4,10,11,15,40–48] Most of the actual measurements have involved elaborate receiving and frequency-measuring equipment—for example, phase-locked receivers capable of determining the phase difference between a satellite signal and one of its harmonics. It is possible to obtain excellent measurements in this way, but at the cost of expensive installations and constant attention by highly skilled technicians. In particular, a phase-locked receiver must be attended continually during a satellite passage lest momentary reductions of signal strength cause the tracking loop to unlock, resulting in loss of the data. Phase-locked receivers are capable of very great sensitivity by virtue of the narrow bandwidths that can be achieved; however, ionosphere beacon satellites to be available in the near future are expected to have adequate power to permit the use of simpler receiving equipment.

It is concluded that the more elaborate techniques of Doppler determination of electron content are more suitable for use at well-staffed and well-equipped laboratories, and that simpler equipment should be relied upon for field station use. It may well happen that Faraday rotation methods prove to be best for field use; however, for the sake of completeness, a method of Doppler analysis with relatively simple equipment, suggested by the work of Ross,[10] will be described here.

In order to determine the ionospheric effect on the Doppler shift, it is necessary to know the Doppler shift that would occur in the absence of the ionosphere. This, in turn, implies an accurate knowledge of the orbit and of the radiated frequency of the satellite. In many cases such knowledge does not exist; hence, it is preferable to determine the free-space Doppler shift by observation. This can be done if the satellite radiates two frequencies: one which is strongly affected by the ionosphere, and another so much higher than the critical frequency corresponding to the maximum electron density that it is substantially free of ionospheric influence. The frequency ratio between the two signals must be known very precisely, so it is convenient that one be a harmonic of the other.

Let the two frequencies be f_1 and f_2. The phase path length corresponding to either is

$$P = \int_r^s n\,ds \ , \qquad (10)$$

where n is the index of refraction for the frequency in question and ds is an element of path length. The number of wave cycles between the satellite at s and the receiver at r is

$$\int_r^s \frac{ds}{\lambda} = \frac{P}{\lambda_0} \ , \qquad (11)$$

in which λ and λ_0 are the wavelengths in the medium and in free space, respectively. The observed frequency is the emitted frequency minus the rate of change of the number of cycles along the path, $f-\dot{P}/\lambda_0$, and the "Doppler shift" is the difference between the observed and emitted frequencies, $\dot{P}/\lambda_0$.

In the ionosphere the index of refraction is given by

$$n^2 = 1 - \frac{80.6\,N}{f^2} \ , \qquad (12)$$

or, if the frequency is well above the critical frequency, by

$$n \approx 1 - \frac{40.3\,N}{f^2} \ . \qquad (13)$$

The phase path length in the ionosphere is thus

$$P_i = \int_r^s \left(1 - \frac{40.3N}{f^2}\right) ds \ , \qquad (14)$$

and in the absence of the ionosphere it is simply

$$P_{fs} = \int_r^s ds \ . \qquad (15)$$

The difference between these quantities, sometimes called the "phase path defect,"[47] is

$$\Delta P = P_{fs} - P_i = \int_r^s (1-n)\,ds \approx \frac{40.3}{f^2}\int_r^s N\,ds \quad . \tag{16}$$

At any point along the ray path

$$ds = \sec i\, dh \quad , \tag{17}$$

where dh is an element of length in the vertical direction and i is the local zenith angle of the ray path. If attention is restricted to those parts of the orbit rather near the zenith and if i_F is the value of i near the maximum of density of the F region, then

$$ds \approx \sec i_F\, dh \quad , \tag{18}$$

and

$$\Delta P \approx \frac{40.3 \sec i_F}{f^2}\int_r^s N\,dh \quad . \tag{19}$$

It happens that the ratio $\Delta P/P_{fs}$ is nearly constant if the orbit is nearly circular and the ionosphere is spherically stratified. Thus

$$\frac{\Delta P}{P_{fs}} = \frac{\Delta \dot{P}}{\dot{P}_{fs}} = \frac{\Delta \ddot{P}}{\ddot{P}_{fs}} \quad , \tag{20}$$

where $\dot{P}$ is the rate of change of frequency and, in particular, $\ddot{P}_{fs}$ is the maximum slope of the free-space Doppler shift curve when plotted as a function of time. In practice, $\ddot{P}_{fs}$ and $\Delta\ddot{P}$ can be evaluated directly from observation, provided that the two frequencies are appropriately chosen from Equations 19 and 20:

$$\int_r^s N\,dh = \frac{f^2 P_{fs} \cos i_F}{40.3}\,\frac{\Delta\ddot{P}}{\ddot{P}_{fs}} \quad , \tag{21}$$

where P_{fs} is simply the geometrical distance from satellite to receiver and can be determined from the known orbital elements, as can i_F; the symbol f refers to the lower of the two radiated frequencies; and $\Delta\ddot{P}$ is determined by comparing the Doppler shifts on the two frequencies.

To illustrate the determination of $\Delta\ddot{P}$ and $\ddot{P}_{fs}$ consider, for example, a satellite radiating phase-locked signals on 40 and 360 Mc. Since ΔP varies inversely as the square of the frequency, the phase path defect will be about 81 times as large on 40 as on 360 Mc. Consider, then, that the 360 Mc wave is not influenced by the ionosphere. A separate superheterodyne radio receiver is used for each frequency (figure 3). An additional "beating oscillator" is provided which has phase-locked outputs near 40 and 360 Mc, and these outputs are injected into the RF input terminals

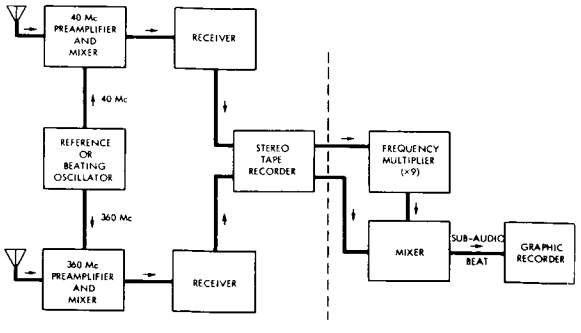


FIGURE 3.—Simplified two-frequency Doppler system.

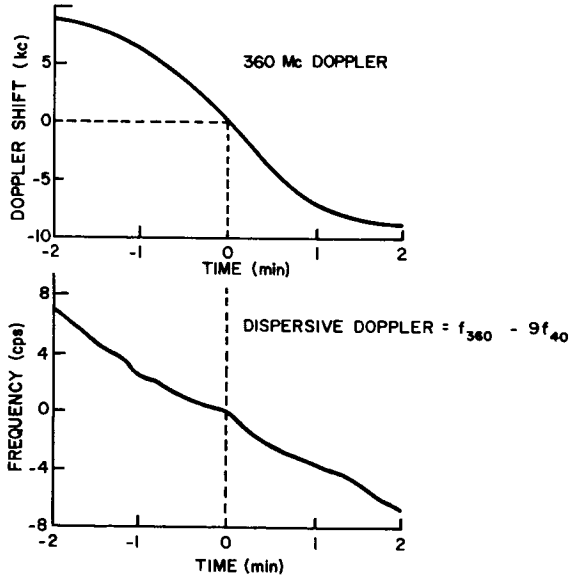


FIGURE 4.—Doppler records.

of the respective receivers. Thus, an AF beat is produced in each receiver, the 360 Mc receiver's beat frequency being *approximately* nine times that of the 40 Mc receiver. Next, the beat note of the 40 Mc receiver is multiplied in frequency by a factor of nine and mixed with the beat note from the 360 Mc receiver, and the "difference frequency" is selected by means of an appropriate filter. This difference frequency is $\Delta\dot{P}/\lambda_0$ and its rate of change is $\Delta\ddot{P}/\lambda_0$. The difference frequency is low enough to be recorded directly on a graphic record. At the same time, the beat frequency from the 360 Mc receiver is recorded separately to permit determination of $\ddot{P}_{fs}$ or $\dot{P}_{fs}$.

The records will appear approximately as shown in figure 4, in which the upper curve is the 360 Mc Doppler shift, approximately equal to $\dot{P}_{fs}$, and the lower curve is the difference frequency or "dispersive Doppler frequency," $\Delta\dot{P}/\lambda_0$. The slope of the upper curve at the inflection point is P_{fs}/λ_0.

Only the two receivers, the beating oscillator, and the tape recorder need be installed at the field station. All equipment to the right of the dashed line in figure 3 may be installed at the central laboratory, and one setup of this kind can be used to transcribe the tape records from several field stations.

The foregoing analysis is the simplest one possible and involves the following assumptions: The satellite is near the zenith and the ionosphere is spherically stratified. The orbit is circular. The high-frequency approximation is satisfied for the lower of the two frequencies, and the higher frequency is so high that its phase velocity is almost that of free space. The effect of the magnetic field is neglected.

THE FARADAY ROTATION OBSERVATION STATION

The Site

It is essential that a quiet site be chosen, as noisy records are apt to be useless. Although the contemplated signal strengths will be more than adequate in most cases, proximity to heavily traveled highways or streets, industrial or commercial activity, or densely settled residential areas will inevitably cause trouble. Automobile ignition systems, farm machinery, neon signs, fluorescent lighting, defective power transformers, and high-voltage power lines are frequent causes of interference at the frequencies in question. Fortunately, noise of these types attenuates rapidly with distance and it has been found that locating the receiving equipment a mile or so from the offending source usually provides sufficient isolation. Thus, a well-isolated site should be chosen.

Terrain does not appear to be particularly important, though reasonably flat ground is desirable. Only if it is desired to erect a radio-interferometer for position measuring purposes is a flat area required, and even then it may be possible to use level metal ground planes erected on sloping terrain.

Antennas

For Faraday rotation measurements the antennas must be plane polarized. Simple half-wavelength dipoles have been used almost universally for this purpose. As the frequencies used are in the HF or the lower VHF bands, a half-wavelength dipole has adequate effective area to yield good signal-to-noise ratios with the signal strengths contemplated. The most valuable Faraday rotation data are received from within, say, 40° of the zenith, whereas data from low-altitude angles are of much less importance. For this reason it is logical that the dipoles be horizontally polarized and situated within a quarter-wavelength of the ground.

For the best quality data, the antenna should have a negligible circular component of polarization, and care should therefore be taken to suppress the effects of antenna currents on the transmission lines. If a coaxial transmission line is used, the antenna should be connected through a carefully adjusted balanced-to-unbalanced transformer. Open-wire lines should be symmetrical and properly matched in impedance. If the transmission lines are of substantial length, consideration should be given to placing preamplifiers in the field near the antennas in order to improve the system noise performance.

It is convenient to have antennas on different frequencies oriented similarly, in order to distinguish between signal strength variations due

to Faraday rotation and those due to satellite motion. Some observers[7] prefer to keep track of the sense of the Faraday rotation; this requires two antennas on the same frequency, with different directional orientation. In general, however, this arrangement does not appear to be essential.

Finally, a word should be said in favor of substantial construction. Experience has shown that a surprisingly large proportion of the technical difficulties experienced by field stations are due to broken antenna or transmission-line wires or to poor connections. The use of sturdy wooden structures and heavy gauge copper-clad steel wires, and proper securing of transmission lines to prevent repeated flexing at connections, are worthwhile precautions. Outdoor splices in coaxial cables should be avoided.

Receivers

Receivers for Faraday-rotation recording are invariably of the superheterodyne type as illustrated in figure 5. As the radio frequencies are above about 20 Mc and the required bandwidths quite narrow, a double-conversion system is usually used. In the past, many high grade communications receivers have been so employed with very good results, including the Hammarlund SP–600, the Collins 51–J, and the military R–390. For best results some modifications are usually required, as will be discussed below. When contemplating a long-term program of routine data gathering, however, it is appropriate to consider receivers specially designed and built for the task. The general purpose receivers usually are very heavy and intricate and thus difficult to maintain, usually have features not required for the task, and often have some undesirable characteristics.

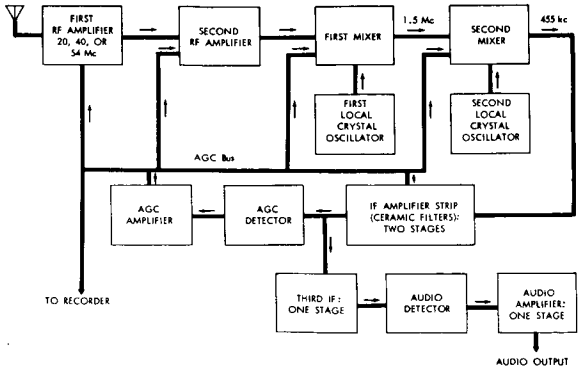


FIGURE 5.—Superheterodyne receiver for Faraday rotation observations.

In order to achieve optimum signal-to-noise ratios it is desirable that the overall bandwidth be as narrow as possible. The limit is set by the Doppler shift of the signal. At 20 Mc this amounts to about 500 cps for a close orbit, requiring a total bandwidth of 1 kc. At 40 Mc the effect is doubled, requiring a bandwidth of 2 kc. The highest frequency likely to be used for Faraday rotation studies is 54 Mc, so that a standard 3-db bandwidth of slightly over 2.5 kc should accommodate all cases. This bandwidth is established in the second IF stage. Ceramic IF filters at 455 kc are now available which require no adjustment and which apparently are drift-free over long periods of time. Combined with transistor amplifiers, these should permit the construction of a virtually maintenance-free IF strip. These can be made very economically in the form of standard modules for ease of replacement.

A low-noise RF amplifier is needed to insure that the sky noise, rather than receiver noise, limits the performance of the system. Two crystal-controlled local oscillators are needed. The first crystal is changed to accommodate satellites of different frequencies. In this way there is little chance that the receiver will be mistuned.

Particular attention should be paid to the automatic gain control circuits. It is desirable to have a logarithmic amplitude response, both to prevent saturation of the receiver or recorder and to sharpen the nulls of the Faraday rotation records. Satellite signals are usually faint, so the AGC circuits should develop appreciable voltage at receiver inputs of the order of 0.2 microvolt. If a commercially built receiver is used it is usually necessary to modify the AGC circuits to remove the "delay bias" in order to accomplish logarithmic response down to these low input levels. The AGC time constant must be small, not higher than 0.1 second for Faraday rotation and 0.02 second for scintillation.

Although not strictly necessary, it is very convenient to have a beat-frequency oscillator in order to permit qualitative checks of the presence of the satellite signal, signal strength, nature and identity of interfering signals, and so on. It is

essential, however, that the BFO should not contribute to the AGC voltage. It is also convenient to have an "s-meter" to permit visual monitoring of the signal strengths, and an audio detector, amplifier, and loudspeaker for use with the BFO.

The following common features of communications receivers are not used: Automatic noise limiter, audio filter or *Q*-multiplier, squelch, crystal IF filter, and general frequency coverage.

Recording Equipment

The recorder is the most expensive and troublesome component of a satellite receiving system. Two general types of recorders are in common use: Magnetic tape recorders and high-speed graphic recorders, each having its advantages and disadvantages for field use in a Faraday rotation program.

The most satisfactory chart recorder from the standpoint of reliability and convenience is the heat-writing variety. In this method a heated stylus marks a wax-impregnated chart, so there are no pen-clogging problems and there is no chance of ink-supply exhaustion. Usually, these instruments produce rectilinear records as well. Their initial costs and operating expenses, unfortunately, are higher than those of the ink-writing recorders.

To obtain the maximum amount of useful data by the Faraday rotation method from a satellite having 20, 40, and 41 Mc signals requires recording of at least three channels of data, excluding timing marks. Most recorders are furnished with an even number of channels; thus, a four channel recorder would be required. The fourth channel is useful as a spare. Timing marks are recorded on the "side pen" or "event marker." If economy were a prime requirement, the 20 Mc channel might be eliminated and a two-channel recorder used, at the sacrifice of the greater scintillation sensitivity and the corrections for refraction and the high-frequency approximation that are obtainable with the 20 Mc signal.

Acceptable specifications for a chart recorder for Faraday rotation work are as follows:

Number of channels: minimum 2, 3 preferred; event marker.
Frequency response: 0 to 50 cps or better.
Chart speed: 5 or 10 mm/sec, preferably both.
Type of recording: Ink-writing or heat-writing.
Transistorized amplifiers desirable for reliability.

Magnetic tape recording can also be used. It has the disadvantages that no visual record is available for immediate inspection, and that it is ultimately necessary to re-record the signals on a graphic recorder. Usually the magnetic tapes from several field stations are sent to a central laboratory where they can be transcribed; thus, only one graphic recorder is needed. Probably the best practice is to transform the variable AGC voltages from the receivers into variable-frequency audio tones by means of voltage-controlled oscillators. Several of these audio tones of different frequencies can be recorded on the same tape channel and separated later by means of appropriate filters. Frequency discriminators are used to convert the tones back to varying dc signals for graphic recording.

For Faraday rotation work the tape recorders need not be of extremely high quality, but should be very reliable in operation. The principal deficiency of an inexpensive tape recorder of the consumer market variety is inconstant tape speed (jitter, wow, flutter) which results in frequency modulation of the record. If not too severe, this should not be a serious defect for purposes of Faraday rotation recording. In any event, it is possible to overcome this problem by recording a constant, standard tone along with the signals, and later electronically subtracting the dc output of the discriminator associated with this tone from the outputs associated with all the other channels. Thus, the graphic recordings should be free from jitter. By proper spacing of the tones in frequency, it should be possible to record five channels of data, including time signals and a speed-calibration tone, on one tape. If a "stereo" recorder is used, one channel can be reserved for voice announcements, time signals, and the speed-calibration tone; and the other channel for the three satellite signals.

By comparison with the graphic recorder the magnetic tape recorder has the following advantages: It is less expensive. A record on tape can be transcribed to a graphic record at any desired speed and any desired number of copies

can be made. Frequently, especially when studying scintillation, it is desirable to have graphic records at more than one speed. If economy is a principal requirement, the magnetic tapes can be reused after they have been transcribed. Magnetic tape recorders are generally more adaptable to unattended operation, though the heat-writing graphic recorders are quite usable in this way. The main disadvantage of the magnetic recorder is the lack of a visual record to assist the operator in making adjustments and corrections while a pass is in progress. As the receiving equipment described in this report is designed for a minimum of operator attention and adjustment, it may well be that this consideration is not applicable.

For an organization planning to operate several field stations the choice of magnetic tape recorders seems a good one, with one graphic recorder and associated filters and discriminators at the central data-handling facility. If only one station is to be operated, however, recording directly on a graphic recorder eliminates the need for the magnetic recorder and associated equipment.

Time Signals

To be of any scientific value, all satellite recordings must contain precise, unambiguous time marks. An error of, say, 0.25 second will generally be quite acceptable, but errors of several seconds will not.

Time marks can be derived from many of the radio time services, and, if the station is so situated that regular time signals can be received reliably, these can be recorded directly from the radio receiver onto the time channel of the magnetic tape. For a graphic recorder, the time ticks can be filtered, rectified, and transformed into low impedance pulses to operate the event-marker pen. These radio time signals are often received by means of ionospheric propagation, however, and are not always reliable. Many of the stations use the same frequencies, and they sometimes interfere. For example, at Adak, Alaska, signals from stations WWVH and JJY interfere so badly that they are useless for automatic registration.

Recently a number of stations have begun transmitting time signals on VLF wavelengths; for example, NBA on 18 kc in the Panama Canal Zone. This station is powerful enough to be heard anywhere on earth. A pocket-sized battery-powered transistorized receiver with a ferrite "loopstick" antenna has been constructed which is capable of receiving these signals sufficiently well for satellite recording, and it has been found that there is essentially no variation of signal strengths with time. Apparently, however, no standard time-signal format has been adopted by this station; and its usefulness is somewhat limited until this is done.

A relatively inexpensive local time-standard can be built by using a tuning fork oscillator and power amplifier to drive a clock motor. Such a device, utilizing a cam-operated switch to provide time ticks every ten seconds with a different identifying tick every minute, demonstrated a precision equivalent to about one second of error per week. Radio time signals can usually be heard with adequate quality to calibrate a local time-standard, even though not good enough for direct recording.

The ideal timing system would be a quartz-crystal oscillator driving a chain of frequency dividers to give hour, minute, and second ticks. This system is the most expensive, but is capable of very high accuracy and can also provide a program function for unattended automatic operation of the satellite recording equipment.

In addition to time marks for the satellite records it is very useful to have a "program clock" to permit scheduling of unattended satellite recording. As was mentioned above, a quartz-crystal oscillator and frequency divider chain can provide for this function, but at relatively high cost. The tuning-fork frequency standard can be used to drive any of a variety of low-cost commercially available program timers that are satisfactory for the purpose. When a program clock is used, it is necessary that unambiguous time marks be placed on the chart in addition to the regular second or minute ticks.

STANDARDS

In order that the results from different stations be comparable, some degree of standardization of observing practice must be achieved. It is always tempting to the observer to experiment with different methods, but it is important to a long-range monitoring program that the records

be taken in a consistent manner throughout the program. Experiments with different methods should be performed independently of the routine data-gathering effort.

Suggested standards for recording satellite data are listed below.

Faraday Rotation and Scintillation

Amplitude scintillation can be studied directly from the records of amplitude variation due to Faraday rotation. The following standards have been found suitable:

Time marks.—Accurate to 0.5 second. A reference should be made at least every minute, preferably every few seconds, and unambiguous time marks should be made at least every five minutes. At least one, and preferably more, of the unambiguous time marks must be available on each record.

Chart speed.—For Faraday rotation studies, chart speeds of 5 to 10 mm per second are generally satisfactory. For passages in the vicinity of noon and at tropical latitudes the higher speed is almost mandatory. In studying scintillation it may sometimes be advantageous to increase the chart speed to 25 or 50 mm per second; however, it is recommended that this be done on an auxiliary recorder so as not to disturb the routine arrangements. For this purpose a magnetic tape recorder is very useful; the tape can be recorded at any convenient speed and can then be transcribed any number of times to paper charts running at a variety of speeds.

Receiving system time-constants.—As was mentioned earlier, the AGC time constant should be not greater than 0.1 second for Faraday rotation recording and not greater than 0.02 second for scintillation. The same limitations apply to other parts of the system, including recorders.

Signal-to-noise ratio.—It is difficult to place a numerical specification on the necessary signal-to-noise ratio for good results. Suffice it to say that the better the quality of the record, the easier is the analysis of the data. With reasonable care it has been found possible to produce records that are essentially noise free; that is, on which the Faraday rotation and scintillation amplitude variations are very well defined, but on which the noise variations are barely discernible. Satellites specifically designed for beacon use will generally have adequate power to produce such records with dipole antennas and receivers such as those described herein.

Amplitude calibration.—No amplitude calibration is necessary for Faraday rotation analysis. For investigation of the geographical and temporal distributions of scintillations, infrequent calibrations may be useful. However, for studies of the detailed structure of the signals and of the irregularities producing scintillation, it is important that frequency calibration of the amplitude-response characteristics be made with a standard signal generator.

Record annotation and logging.—One of the most persistent difficulties in the collection of high-quality satellite records is the failure of operators properly to annotate the records. Each record should bear the following information:

Date.
Time of beginning and ending of pass.
Unambiguous and precise identification of one or more of the periodic time marks on the chart.
Time system used (GMT or zone time).
Antenna identification for each trace.
Satellite identification.
Frequencies for each trace.
Identification of noise bursts or interfering signals.
Station name and/or location.
Operator's name.

In addition, a logbook should be kept, containing the above information for each pass and also notations concerning equipment maintenance or modification, spurious signals or noise on the records, and any unusual events or phenomena observed.

Doppler Method

Most of the foregoing comments also apply to the Doppler method. As the dispersive Doppler records made in the field will almost certainly be recorded on magnetic tape, no definite standard need be adopted for tape speed. In the laboratory, when the records are transcribed to paper charts, the chart speed can be adjusted to allow easy counting of the dispersive Doppler cycles.

Cost of Equipping a Field Station

It is estimated that the ground-based instrumentation required to perform the geophysical mission described herein would cost approximately $3,200.

ACKNOWLEDGMENTS

The author gratefully acknowledges the valuable contributions to this paper made by Dr. W. J. Ross of the Pennsylvania State University, Dr. K. C. Yeh of the University of Illinois, and Dr. O. K. Garriott of Stanford University.

REFERENCES

1. Al'pert, Ya. L., "Results of Investigations of the Outer Region of the Ionosphere According to Observations of Radio Signals of the First Artificial Earth Satellite," New York: U.S. Joint Publications Research Service, 1958 (JPRS/NY Rept. No. 724); Translation from *Preliminary Results of Scientific Research by Means of the First Soviet Earth Satellites and Rockets*, No. 1, Part II, pp. 40–108 (in Russian).
2. Al'pert, Ya. L., Dobryakova, F. F., et al., "Electron Concentration in the Upper Ionosphere as Determined from Radio Signals of the First Earth Satellite," *Soviet Physics—Doklady*, **3**(3): 584–587, January 1959; Translation from *Doklady Akademii Nauk SSSR* **120**(4): 743–746, 1958.
3. Blackband, W. T., and Burgess, B., et al., "Deduction of Ionospheric Electron Content from the Faraday Fading of Signals from Artificial Earth Satellites," *Nature*, **183**(4669): 1172–1174, April 25, 1959.
4. Hibberd, F. H., and Thomas, J. A., "The Determination of the Electron Distribution in the Upper Ionosphere from Satellite Doppler Observations," *J. Atmos. Terrest. Phys.*, **17**(1/2): 71–81, December 1959.
5. Vassy, E. J., "Densité Électronique au-dessus de F-maximum, Déduite de l'Émission des Satellites," in: *The Upper Atmosphere above F2-Maximum*, ed. by H. Poeverlein, North Atlantic Treaty Organization, Advisory Group for Aeronautical Research and Development, AGARDograph 42, May 1959, pp. 263–270.
6. Blackband, W. T., "The Determination of Ionospheric Electron Content by Observation of Faraday Fading," *J. Geophys. Res.*, **65**(7): 1987–1992, July 1960.
7. Garriott, O. K., "The Determination of Ionospheric Electron Content and Distribution from Satellite Observations, I and II," *J. Geophys. Res.*, **65**(4): 1139–1157, April 1960.
8. Hame, T. G., and Stuart, W. D., "The Electron Content and Distribution in the Ionosphere," *Proc. IRE*, **48**(10): 1786–1787, October 1960.
9. Little, C. G., and Lawrence, R. S., "The Use of Polarization Fading of Satellite Signals to Study the Electron Content and Irregularities in the Ionosphere," in: *Space Research, Proc. 1st Internat. Space Sci. Sympos. Nice, January* 1960, ed. by H. K. Bijl, Amsterdam: North-Holland Pub. Co., 1960, pp. 340–365.
10. Ross, W. J., "The Determination of Ionospheric Electron Content from Satellite Doppler Measurements, 1 and 2," *J. Geophys. Res.*, **65**(9): 2601–2615, September 1960.
11. de Mendonca, F., "Ionospheric Electron Content and Variations Measured by Doppler Shifts in Satellite Transmissions," Stanford Univ., Radioscience Lab. (Preprint submitted with Semiannual Status Report for the Period July 20, 1961, to January 20, 1962, NASA Grant NsG–30–60), January 1962.
12. Yeh, K. C., and Swenson, G. W., Jr., "Ionospheric Electron Content and Its Variations Deduced from Satellite Observations," *J. Geophys. Res.*, **66**(4): 1061–1067, April 1961.
13. Munro, G. H., "Diurnal Variations in the Ionosphere Deduced from Satellite Radio Signals," *J. Geophys. Res.*, **67**(1): 147–156, January 1962.
14. Slee, O. B., "Radio Scintillations of Satellite 1958 α," *Nature*, **181**(4623): 1610–1612, June 7, 1958.
15. Aitchison, G. J., Thomson, J. H., and Weekes, K., "Some Deductions of Ionospheric Information from the Observations of Emissions from Satellite 1957 $\alpha 2$," *J. Atmos. Terrest. Phys.*, **14**(3/4): 236–248, 1959.
16. Kent, G. S., "High Frequency Fading Observed on the 40 Mc/s Wave Radiated from Artificial Satellite 1957α," *J. Atmos. Terrest. Phys.*, **16**(1/2): 10–20, October 1959.
17. Parthasarathy, R., and Reid, G. C., "Signal Strength Recordings of the Satellite 1958 $\delta 2$ (Sputnik III) at College, Alaska," *Proc. IRE*, **49**(1): 78–79, January 1959.
18. Yeh, K. C., and Swenson, G. W., Jr., "The Scintillation of Radio Signals from Satellites," *J. Geophys. Res.*, **64**(12): 2281–2286, December 1959.
19. de Mendonca, F., Villard, O. G., Jr., and Garriott, O. K., "Some Characteristics of the Signal Received from 1958 δ^2," *Proc. IRE*, **48**(12): 2028–2030, December 1960.
20. Frihagen, J., and Tröim, J., "Scintillations of the 20 Mc/s Signals from the Earth Satellite 1958 δII," *J. Atmos. Terrest. Phys.*, **18**(1): 75–78, April 1960.
21. Mawdsley, J., "Fading of Satellite Transmissions and Ionospheric Irregularities," *J. Atmos. Terrest. Phys.*, **18**(4): 344, August 1960.
22. Bain, W. C., "Irregular Fading of Satellite Transmissions," *Nature*, **189**(4759): 129, January 14, 1961.
23. Frihagen, J., and Tröim, J., "On the Large Scale Regions of Irregularities Producing Scintillation of Signals Transmitted from Earth Satellites," *J. Atmos. Terrest. Phys.*, **20**(2/3): 215–216, March 1961.

24. KENT, G. S., "High Frequency Fading of the 108 Mc/s Wave Radiated from an Artificial Earth Satellite as Observed at an Equatorial Station," *J. Atmos. Terrest. Phys.*, **22**(2): 255–269, December 1961.
25. MAWDSLEY, J., and RICHARDS, I. R., "Ionospheric Scattering of Satellite Transmissions," *Nature*, **189**(4768): 906–907, March 18, 1961.
26. SINGLETON, D. G., LYNCH, G. J. E., and THOMAS, J. A., "Field-Aligned Ionospheric Irregularities and the Scintillation of Satellite Radio Transmissions," *Nature*, **189**(4758): 30–31, January 7, 1961.
27. SWENSON, G. W., Jr., and YEH, K. C., "Summary of Satellite Scintillation Observations at the University of Illinois," Univ. of Illinois, Engng. Exp. Sta., Elect. Engng. Res. Lab., Report on NASA Grant NsG–24–59, August 4, 1961.
28. DEWAN, E., "An Interesting Propagation Effect of Sputnik I," U.S. Air Force Cambridge Res. Ctr., Electromagnetic Radiation Lab. AFCRC–TR–58–354, December 1958.
29. GARRIOTT, O. K., and VILLARD, O. G., Jr., "Antipodal Reception of Sputnik III," *Proc. IRE*, **46**(12): 1950, December 1958.
30. WELLS, H. W., "Unusual Propagation at 40 Mc from the USSR Satellite," *Proc. IRE*, **46**(3): 610 March 1958.
31. DEWAN, E. M., "Unusual Propagation of Satellite Signals," *Proc. IRE*, **47**(11): 2020, November 1959.
32. WOYK, E. (E. CHVOJKOVÁ), "The Antipodal Reception of Sputnik III," *Proc. IRE*, **47**(6): 1144, June 1959.
33. WHITNEY, H., STRICK, H., et al., "Sudden Amplitude Variations of Sputnik III Signals," *J. Geophys. Res.*, **65**(12): 4210–4212, December 1960.
34. CHVOJKOVÁ, E. (E. WOYK), "Antipodal Reception of Satellite Signals and Telecommunication on Small Planets," *Bull. Astronom. Institute Czechoslovak Acad. Sci.*, **12**(1): 1–7, 1961.
35. YEH, K. C., and SWENSON, G. W., Jr., "Observed Field Strength in the Neighborhood of the Skip Distance," *J. Geophys. Res.*, **66**(2): 654–656, February 1961.
36. BOWHILL, S. A., "The Faraday-Rotation Rate of a Satellite Radio Signal," *J. Atmos. Terrest. Phys.*, **13**(1/2): 175–176, 1958.
37. BROWNE, I. C., and EVANS, J. V., et al., "Radio Echoes from the Moon," *Proc. Phys. Soc. (London)*, **69**B(9): 901–920, September 1956.
38. YEH, K. C., and GONZALEZ, V. H., "Note on the Geometry of the Earth Magnetic Field Useful to Faraday Effect Experiments," *J. Geophys. Res.*, **65**(10): 3209–3214, October 1960.
39. YEH, K. C., "Second-Order Faraday Rotation Formulas," *J. Geophys. Res.*, **65**(8): 2548–2550, August 1960.
40. HIBBERD, F. H., "The Effect of the Ionosphere on the Doppler Shift of Radio Signals from an Artificial Satellite," *J. Atmos. Terrest. Phys.*, **12**(4): 338–340, 1958.
41. WEEKES, K., "On the Interpretation of the Doppler Effect from Senders in an Artificial Satellite," *J. Atmos. Terrest. Phys.*, **12**(4): 335–338, 1958.
42. BERNING, W. W., "Earth Satellite Observations of the Ionosphere," *Proc. IRE*, **47**(2): 280–288, February 1959.
43. HUTCHINSON, H. P., "Application of Satellite Doppler Shift Measurements. Part II. Slant Range at Nearest Approach," in: *IRE National Convention Record*, New York: Institute of Radio Engineers, 1959, Vol. 7, Part 5, pp. 61–66.
44. KELSO, J. M., "Doppler Shifts and Faraday Rotation of Radio Signals in a Time-Varying, Inhomogeneous Ionosphere. Part I. Single Signal Case," *J. Geophys. Res.*, **65**(12): 3909–3914, December 1960.
45. DE MENDONCA, F., and GARRIOTT, O. K., "The Effect of the Earth's Magnetic Field on Measurements of the Doppler Shift of Satellite Radio Transmissions," Stanford Univ., Radioscience Lab. (Preprint submitted with Semiannual Status Report for the Period July 20, 1961, to January 20, 1962, NASA Grant NsG–30–60), December 1961.
46. GARRIOTT, O. K., and NICHOL, A. W., "Ionospheric Information Deduced from the Doppler Shifts of Harmonic Frequencies from Earth Satellites," *J. Atmos. Terrest. Phys.*, **22**(1): 50–63, September 1961.
47. GARRIOTT, O. K., and BRACEWELL, R. N., "Satellite Studies of the Ionization in Space by Radio," in: *Advances in Geophysics*, New York: Academic Press, 1961, Vol. 8, pp. 85–135.
48. HENDERSON, R. E., "Measuring the Doppler Frequency Shift on Satellite Transmissions," *Brit. Commun. and Electronics*, **8**(7): 506–512, July 1961.

ADDITIONAL BIBLIOGRAPHY

1956

DANIELS, F. B., "Electromagnetic Propagation Studies with a Satellite Vehicle," in: *Scientific Uses of Earth Satellites*, ed. by J. A. Van Allen, Ann Arbor: University of Michigan Press, 1956, pp. 276–282.
HAGEN, J. P., "The Exploration of Outer Space with an Earth Satellite," *Proc. IRE*, **44**(6): 744–747, June 1956.
MENGEL, J. T., "Tracking the Earth Satellite, and Data Transmission, by Radio," *Proc. IRE*, **44**(6): 755–760, June 1956.
VAN ALLEN, J. A., "The Scientific Value of the Earth Satellite Program," *Proc. IRE*, **44**(6): 764–767, June 1956.

1957

BROWN, R. R., and GREEN, P. E., Jr., et al., "Radio Observations of the Russian Earth Satellite," *Proc. IRE*, **45**(11): 1552–1553, November 1957.

PETERSON, A. M., and Staff, "Radio and Radar Tracking of the Russian Earth Satellite," *Proc. IRE*, **45**(11): 1553–1555, November 1957.

Staff of the Royal Aircraft Establishment, Farnborough, "Obesrvation on the Orbit of the First Russian Earth Satellite," *Nature*, **180**(4593): 937–941, November 9, 1957.

1958

BRACEWELL, R. N., and GARRIOTT, O. K., "Rotation of Artificial Earth Satellites," *Nature*, **182**(4638): 760–762, September 20, 1958.

BRIGGS, B. H., "A Study of the Ionospheric Irregularities which Cause Spread-F Echoes and Scintillations of Radio Stars," *J. Atmos. Terrest. Phys.*, **12**(1): 34–45, 1958.

KRAUS, J. D., and ALBUS, J. S., "A Note on Some Signal Characteristics of Sputnik I," *Proc. IRE*, **46**(3): 610–611, March 1958.

KRAUS, J. D., and DREESE, E. E., "Sputnik I's Last Days in Orbit," *Proc. IRE*, **46**(9): 1580–1587, September 1958.

KRAUS, L., and WATSON, K. M., "Plasma Motions Induced by Satellites in the Ionosphere," *Phys. of Fluids*, **1**(6): 480–488, November–December, 1958.

MUNRO, G. H., and WHITE, R. B., "Observations in Australia of Radio Transmissions from the First Artificial Earth Satellite," *Nature*, **181**(4602): 104, January 11, 1958.

WARWICK, J. W., "Radio Observations of Sovict Satellites 1957 Alpha 2 and 1957 Beta 1," *IGY Satellite Rept. Series*, No. 5, June 30, 1958.

WOYK, E., (E. CHVOJKOVÁ), "Investigation of the Ionosphere Using Signals from Earth Satellites," *Nature*, 182(4646): 1362–1363, November 15, 1958.

1959

ARENDT, P. R., "On the Existence of a Strong Magneto-Ionic Effect Topside of the F Maximum of the Kenelly-Heaviside Layer," *J. Appl. Phys.*, **30**(5): 793–795, May 1959.

BRITO, J. M., "A Correction Necessary for the Application of the Doppler Effect to the Measurements of Distances to Satellites," *Proc. IRE*, **47**(11): 2023, November 1959.

DANIELS, F. B., and BAUER, S. J., "The Ionospheric Faraday Effect and Its Applications," *J. Franklin Inst.*, **267**(3): 187–200, March 1959.

HAME, T. G., and KENNAUGH, E. M., "Recordings of Transmissions from the Satellite 1958 Δ_2 at the Antenna Laboratory, The Ohio State University," *Proc. IRE*, **47**(5): 991–992, May 1959.

HIBBERD, F. H., "The Faraday Fading of Radio Waves from an Artificial Satellite," *J. Geophys. Res.*, **64**(8): 945–948, August 1959.

KALLMANN, H. K., "A Preliminary Model Atmosphere Based on Rocket and Satellite Data," *J. Geophys. Res.*, **64**(6): 615–623, June 1959.

KAZANTSEV, A. N., "Absorption and Electron Distribution in the F_2 Layer Determined from Measurements of Transmitted Radio Signals from Earth Satellites," *Planet. Space Sci.*, **1**(2): 130–135, April 1959.

KRASSOVSKY, V. I., "Exploration of the Upper Atmosphere with the Help of the Third Soviet Sputnik," *Proc. IRE*, **47**(2): 289–296, February 1959.

MUNRO, G. H., and HEISLER, L. H., "Recording Radio Signals from Earth Satellites," *Nature*, **183**(4664): 809–810, March 21, 1959.

PAETZOLD, H. K., "Observations of the Russian Satellites and the Structure of the Outer Terrestrial Atmosphere," *Planet. Space Sci.*, **1**(2): 115–124, April 1959.

PARTHASARATHY, R., BASLER, R. P., and DEWITT, R. N., "A New Method for Studying the Auroral Ionosphere Using Earth Satellites," *Proc. IRE*, **47**(9): 1660, September 1959.

SCHILLING, G. F., and STERNE, T. E., "Densities and Temperatures of the Upper Atmosphere Inferred from Satellite Observations," *J. Geophys. Res.*, **64**(1): 1–4, January 1959.

STOREY, L. R. O., "A Method for Measuring Local Electron Density from an Artificial Satellite," *J. Res. Nat. Bur. Standards*, **63**D(3): 325–340, November–December 1959.

WARWICK, J. W., "Decay of Spin in Sputnik I," *Planet. Space Sci.*, **1**(1): 43–49, January 1959.

1960

AARONS, J., WHITNEY, H. E., et al., "Armospheric Phenomena Noted in Simultaneous Observations of 1958 δII (Sputnik III)," U.S. Air Force, Electronics Research Directorate ERD–TR–60–174, August 1960; also in *Planetary and Space Sci.*, **5**(3): 169–184, July 1961.

BIJL, H. K. (ed.), "Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January 1960," Amsterdam: North-Holland Pub. Co., 1960.

GARIOTT, O. K., and LITTLE, C. G., "The Use of Geostationary Satellites for the Study of Ionospheric Electron Content and Ionospheric Radio-Wave Propagation," *J. Geophys. Res.*, **65**(7): 2025–2027, July 1960.

GRAVES, C. D., "Radio Propagation Measurements Using the Explorer VI Satellite," *J. Geophys. Res.*, **65**(9): 2585–2587, September 1960.

KLINKER, L., SCHMELOVSKY, K. H., and KNUTH, R., "Jahreszeitliche Variationen der Mittleren Elektronenkonzentration Zwischen 400 und 1200 Km Höhe," *Naturwissenschaften*, **47**(9): 197–198, May 1960.

KRAUS, J. D., HIGGY, R. C., and CRONE, W. R., "The Satellite Ionization Phenomenon," *Proc. IRE*, **48**(4): 672–678, April 1960.

LOCKWOOD, G. E. K., "Spin Rate of the Satellite Echo I as Determined by a Tracking Radar," *Canadian J. Phys.* **38**(12): 1713, December 1960.

MACRAKIS, M. S., "A Possible Long-Range Communications Link Between Ground and Low-Orbiting Satellites," *J. Atmos. Terrest. Phys.*, **19**(3/4): 260–271, December 1960.

MUNRO, G. H., "Ionospheric Information from Satellite Signals," *Nature*, **187**(4742): 1017–1018, September 17, 1960.

RAND, S., "Wake of a Satellite Traversing the Ionosphere," *Phys. of Fluids*, **3**(2): 265–273, March–April 1960.

1961

AL'PERT, YA. L., "Investigation of the Ionosphere and of the Interplanetary Gas With the Aid of Artificial Satellites and Space Rockets," *Soviet Physics—Uspekhi*, **3**(4): 479–503, January–February 1961; translated from *Uspekhi Fizicheskikh Nauk*, **71**(3): 369–409, July 1960.

ANDERSON, L. J., "108–216 Mc/s Radio Signals from Satellites below the Horizon," *Nature*, **190**(4777): 708, May 20, 1961.

ANDERSON, R. E., "Sideband Correlation of Lunar and Echo Satellite Reflection Signals in the 900-Mc Range," *Proc. IRE*, **49**(6): 1081–1082, June 1961.

CAIN, J. C., SHAPIRO, I. R., et al., "A Note on Whistlers Observed Above the Ionosphere," *J. Geophys. Res.*, **66**(9): 2677–2680, September 1961.

CAPON, I. N., "The Application of Ray Tracing Methods to Radio Signals From Satellites," *Proc. Phys. Soc. (London)*, **77**(2): 337–345, February 1, 1961.

CROFT, T. A., and VILLARD, O. G., Jr., "An HF Radar Search for Possible Effects of Earth Satellites Upon the Upper Atmosphere," *J. Geophys. Res.*, **66**(10): 3109–3118, October 1961.

CROFT, T. A., "An HF Radar Search for the Effects of Earth Satellites Upon the Ionosphere," Stanford Univ., Radioscience Lab. Technical Rept. No. 24, March 10, 1961.

HOLLAND, A. C., "The Effects of Atmospheric Refraction on Angles Measured From a Satellite," *J. Geophys. Res.*, **66**(12): 4171–4175, December 1961.

JAKES, W. C., "A Transatlantic Communication Experiment via Echo I Satellite," *Nature*, **190**(4777): 709, May 20, 1961.

KALLMANN-BIJL, H. K., "Daytime and Nighttime Atmospheric Properties Derived From Rocket and Satellite Observations," *J. Geophys. Res.*, **66**(3): 787–795, March 1961.

MEDVED, D. B., "Secondary Electron Emission and the Satellite Ionization Phenomenon," *Proc. IRE*, **49**(6): 1077–1078, June 1961.

SØRENSEN, E. V., "Magneto-Ionic Faraday Rotation of the Radio Signals on 40 Mc From Satellite 1957 α (Sputnik I)," *IRE Trans. on Antennas and Propagation*, AP-**9**(3): 241–247, May 1961.

TITHERIDGE, J. E., "The Calculation of the Electron Density in the Ionosphere From Elevation-Angle Measurements on Artificial Satellites," *J. Geophys. Res.*, **66**(10): 3103–3107, October 1961.

TOMAN, K., "Minimum Range to Artificial Earth Satellites," *Nature*, **190**(4773): 333–334, April 22, 1961.

YEH, K. C., "Propagation of Spherical Waves in a Weak Random Medium," Univ. of Illinois, Engng. Exp. Sta., Elect. Engng. Res. Lab., Report on NASA Grant NsG–24–59, December 1961.

YEH, K. C., and CHOW, H., "Variations of Ionospheric Electron Content During Disturbances," Univ. of Illinois, Engng. Exp. Sta., Elect. Engng. Res. Lab., Report on NASA Grant NsG–24–59, August 1, 1961.

DIRECT MEASUREMENTS OF HELIUM AND HYDROGEN ION CONCENTRATION AND TOTAL ION DENSITY TO AN ALTITUDE OF 940 KILOMETERS

H. A. TAYLOR, JR., L. H. BRACE, H. C. BRINTON, AND C. R. SMITH

Goddard Space Flight Center

A Bennett-type ion mass spectrometer and a cylindrical electrostatic probe were carried to 940 km above Wallops Island, Va., at midday in October 1961. The raw data from each unit (ion currents versus altitude) are presented and are interpreted in terms of helium, hydrogen, oxygen, and total ion concentration. The He^+ and H^+ concentrations derived from the ion mass spectrometer experiment were found to rise steeply above 350 km, He^+ reaching a maximum value of 7×10^3/cc at 575 km and H^+ increasing monotonically with altitude, ultimately reaching about 1.5×10^3/cc near apogee. These data are used with the O^+ concentrations derived from the electrostatic probe data to calculate the atmospheric temperature, the mean ion mass, and the total ion concentration as functions of altitude. The temperature of the upper F region was found to be $1235°\pm40°$K. The mean ion mass decreased from 16 AMU at 400 km to 12 AMU at apogee where it was decreasing rapidly. The measured ion distributions and those predicted by the theories of hydrostatic diffusion and ion-atom interchange are compared.

INTRODUCTION

At 1740 GMT (1240 local time) on October 10, 1961, an Argo–D4 rocket (NASA 8.23) was launched eastward into the ionosphere above Wallops Island, Va., carrying a Bennett-type ion mass spectrometer and a cylindrical electrostatic probe to an altitude of 940 km and a range of 1200 km. The mass range of the spectrometer had been selected to permit optimum resolution of both hydrogen and helium ions which had been postulated for the upper F region by Nicolet (1961) and have since been inferred from rocket and satellite measurements (Hanson, 1962; Bourdeau et al., 1962). The successful operation of the spectrometer permitted the first high resolution direct measurements of the H^+ and He^+ concentrations in the exosphere. The electrostatic probe was included in the payload to permit a measurement of the total ion density and the electron temperature.

The equipment was placed aboard the vehicle as shown in figure 1, the ion spectrometer looking forward during rocket ascent and the cylindrical probe protruding from the side of the payload. In flight, the rocket spun at a rate of 1 rps about its longitudinal axis with a precession cone half-angle less than 1°. The spin axis, essentially in the plane of the trajectory, made an angle of 15° with the vertical. During the flight approximately 600 spectrums and 5000 probe volt-ampere characteristics were recorded.

In addition to the spectrometer and probe, an optical aspect sensor was included in the payload to permit determination of sensor orientation; this information aided in the conversion of spectrometer currents to ambient ion concentrations. Ionosonde measurements were recorded at Wallops Island throughout the flight to provide electron density data up to the F_2 maximum.

THE ION SPECTROMETER

The Bennett radio-frequency spectrometer was similar in design to that flown at Wallops Island by Taylor and Brinton (1961). However, the

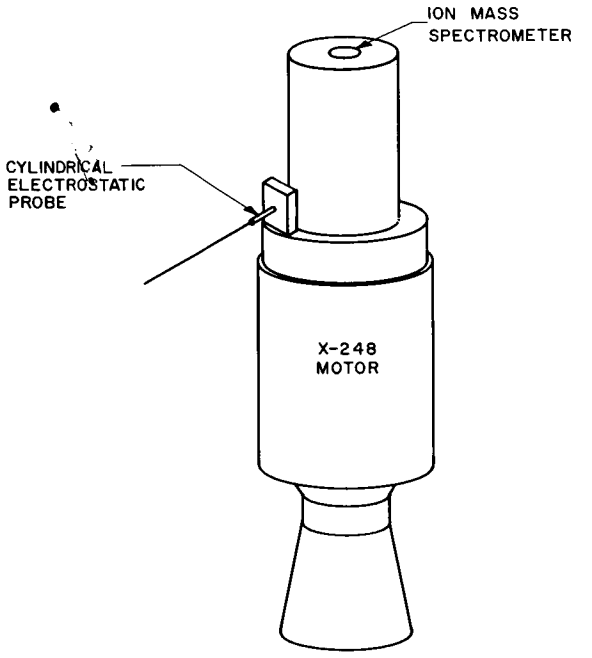


FIGURE 1.—Locations of equipment on NASA 8.23.

mass range and sensitivity were modified to permit unambiguous detection of ions in the mass range of 1–5 AMU. The theory of operation of the spectrometer together with some of the technical details of this experiment has been presented previously (Taylor et al., 1962). Although the theory is not elaborated upon here, it is important to consider several basic parameters of the measurement and their influence on the interpretation of the data.

The basic relationship between resonant mass, tube geometry, and potentials applied to the spectrometer grids is given by the equation

$$M = 0.266(|V_A| + |V_B|)/S^2F^2 \qquad (1)$$

where M is the resonant mass, V_A is the instantaneous value of the negative sawtooth accelerating potential, V_B is a positive bias potential, S is the spacing between grids in the RF analyzer stages, and F is the frequency of the RF potential applied to the analyzer sections. For this flight a three-stage, 5–3 cycle tube was used, with $V_B = 20$ volts dc, $S = 0.75$ cm, $F = 5$ Mc/s, and $V_A = 0$ to -270 volts dc. This provided a mass range of approximately 1 to 5 AMU with a resolution of 0.5 AMU. This spectrum was analyzed at the sweep rate of V_A, which was 1 cps.

An important parameter not contained in (1) is the stopping potential V_S, which regulates the number of ions reaching the collector. In this process V_S controls the efficiency of the spectrometer and also determines the ratio of fundamental to harmonic ion content. The effectiveness of V_S is dependent on the energy gained by ions in traversing the spectrometer and is thus also dependent on energy that ambient ions may gain owing to the effects of external potentials and vehicle velocity.

To insure adequate efficiency and sensitivity to harmonic ions, the V_S level was stepped automatically through an appropriate range of voltage. The harmonic ions detected were analyzed by a technique similar to that developed by Johnson (1958) to confirm the identification of the fundamental ions as 1 AMU and 4 AMU. Although there was some evidence in the data of an apparent lowering of the effective value of V_S that has not as yet been accounted for in terms of measured experimental parameters, the relationships between fundamental and harmonic ions observed in the flight data agree favorably with both theoretical predictions and laboratory calibrations.

A photograph of a segment of the flight data, recorded near apogee, is shown in figure 2. Only two fundamental ion current peaks, H^+ and He^+, were observed during the flight. The raw ion currents I_c measured at the collector of the spectrometer during ascent are plotted in figure 3.

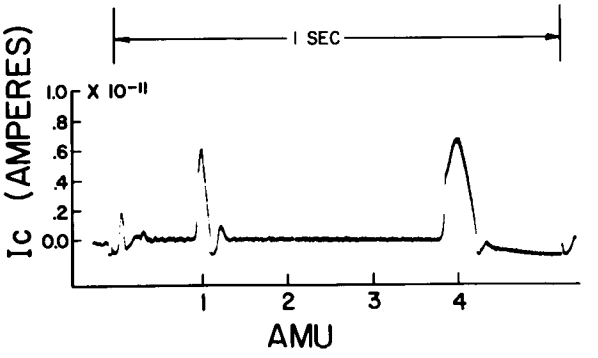


FIGURE 2.—A sample positive ion spectrum recorded near apogee on NASA 8.23.

As is indicated by the error bars, the spread, or deviation, in the raw current data points is slight, as might be expected because of the high degree of aspect stability experienced during flight.

Ion currents collected during descent when the spectrometer was exposed to the wake of the payload are not presented. Although certain parts of the descent data compare favorably with the ascent data, the difficulties inherent in the interpretation of the wake effect preclude the inclusion of the descent data.

Conversion of ion currents to densities requires the determination of over-all experimental sensitivity, which is a function of vehicle potential and velocity, plasma sheath configuration adjacent to the orifice, spectrometer efficiency, and amplifier sensitivity. Both the limiting amplifier sensitivity and spectrometer efficiency were measured directly in the laboratory. These measured values, together with the best estimates of the other effects, have been combined to determine a maximum sensitivity of approximately 10 ions/cc. The relationships of those parameters which were not subject to direct measurement, together with the resulting technique for determination of ion density, are discussed in detail in Appendix 1.

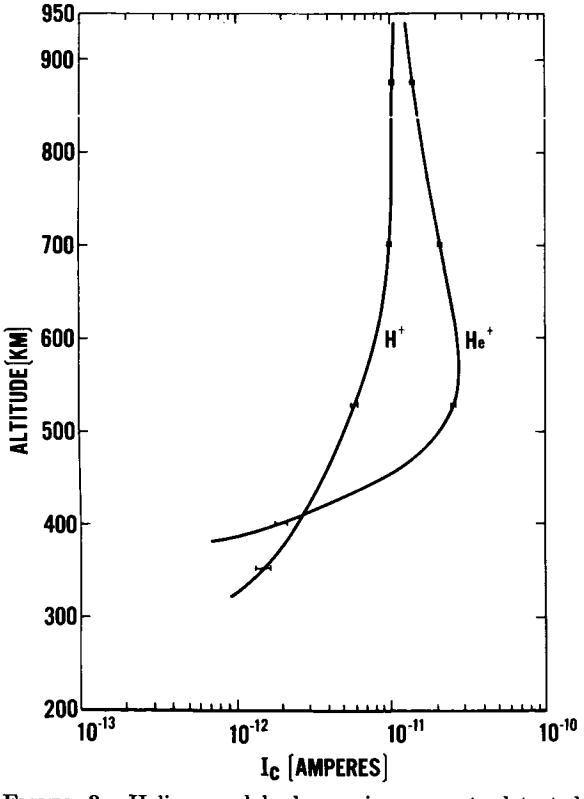


FIGURE 3.—Helium and hydrogen ion currents detected by the ion mass spectrometer.

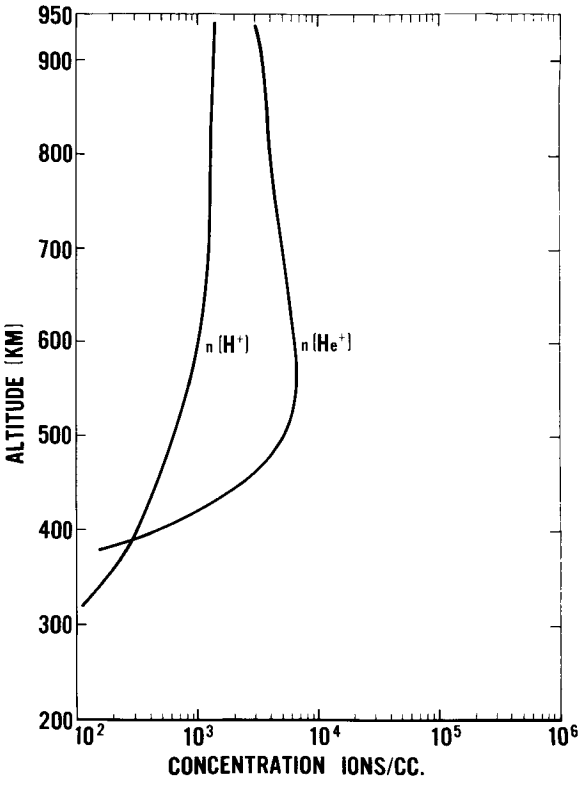


FIGURE 4.—Helium and hydrogen ion concentration profiles derived from the ion mass spectrometer currents shown in figure 3.

It is interesting to consider what observations can be made from the H^+ and He^+ current distribution curves if a simplified conversion to ion density is performed. The part of the trajectory where the perturbing influences of velocity and plasma sheath are most easily interpreted is at peak altitude. At that point the densities attributed to H^+ and He^+ are estimated to be 1.4×10^3 ions/cc and 3.2×10^3 ions/cc, respectively. Figure 4 is a plot of ascent density distributions obtained from the current distributions as discussed in Appendix 1. Although the current curves are now plotted as densities, the distributions remain identical to the current profiles of figure 3, since no modifying functions have been applied to the original curves.

As is discussed in Appendix 1, it is believed that the most accurate conversion of collector

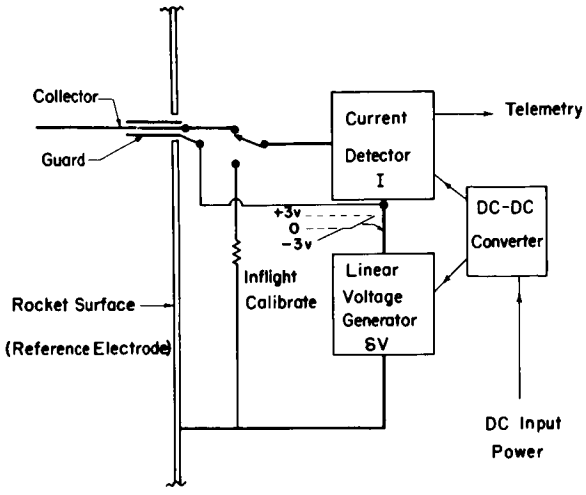


FIGURE 5.—Block diagram of the electrostatic probe.

current I_c to the number densities $n(H^+)$ and $n(He^+)$ may ultimately require a simultaneous solution of the effects of velocity, sheath, and effective orifice area, which, to date, have yielded only to separate analyses. Although it is recognized that these unsolved relationships may modify the shape of the derived composition, it is believed that the attracting field of the orifice grid was the predominating influence, and that the measured current distributions closely represent the ambient ion distributions. With these qualifications then, the H^+ and He^+ concentration profiles will be compared in later sections with the density data obtained by the electrostatic probe.

THE ELECTROSTATIC PROBE

The cylindrical probe used in this flight was nearly identical to one described by Spencer et al. (1962) which was carried as a secondary experiment on two flights of a series of ejected instrument packages launched into the F region. Figure 5 shows a block diagram of the measurement system employed. The metallic payload mounting and instrument housings acted collectively as a plasma reference electrode from which a sawtooth voltage was applied to both the guard and collector elements of the probe sensor. Figure 6 is a photograph of a one-second section of telemetry record showing a series of volt-ampere characteristics recorded at 510 km during ascent. In the central part of the photograph, the ionosphere measurements are interrupted by a brief period of inflight current-calibration during which the probe was electrically disconnected and a known resistance substituted into the measurement circuit. The resulting current wave form was used to calibrate the current channel. Diode limiting at the output of the detector was used to prevent negative currents from driving the telemetry out of band. At the left and right of the calibration wave forms, the positive parts of a series of cylinder current characteristics are displayed. These represent the net current ($I_{net} = I_p - I_e$, the sum of the ion and electron currents) that was collected as the applied voltage was swept through the range of +2 volts to −3 volts, pausing briefly at zero volts for purposes of reference. As is shown in the calibration, the sensitivity of this particular channel was selected to resolve only net positive currents. Note that at zero applied voltage a net positive current flows to the probe (indicating that the collector was being held more negative than its equilibrium potential). As the applied voltage was swept negative, the ambient electrons were entirely rejected and the probe current was due to positive ions alone. Conversely, for positive applied voltages the electron current quickly exceeded the ion current, thus driving the net current negative and out of the range of this detector. A second detector, having the correct sensitivity to resolve the much larger electron currents, was to be used periodically, but a programmer failure prevented its use and permitted the ion current detector to operate full time. This improved the resolution of the ion current data but hampered the analysis of the current characteristics for electron temperature.

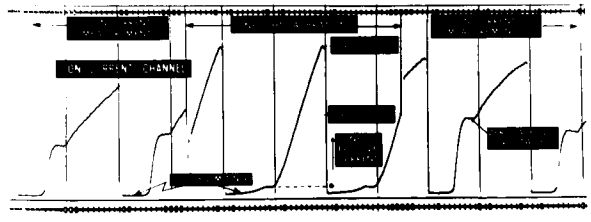


FIGURE 6.—Photograph of a segment of telemetry record showing electrostatic probe ion currents measured at 510 km, and an inflight current calibration.

Figure 7 shows a running mean of the maximum ion currents (measured at maximum negative voltage) recorded throughout the flight. The times of F_2 passage are evident in the current maximums at 1743 and 1757 GMT on ascent and descent, respectively. It should be noted that these positive currents result not only from the collection of positive ions but include a small component due to photoemission; i.e., ejection of electrons from the collecting surface as the result of solar extreme ultraviolet radiation and X-ray impingement. The photocurrent has been evaluated by means of the small jump-discontinuity it produces in the current as the rocket spin carries the probe through the shadow of the payload structure. At apogee, where the total positive current had fallen to 8×10^{-9} ampere, the photocurrent was found to be 2×10^{-9} ampere. This value is in acceptable agreement with rocket measurements of Hinteregger et al. (1959) of photoemission which, when translated to the geometry and area of the cylindrical collector, correspond to a value of about 3.5×10^{-9} ampere. The decrease in solar activity since the time of Hinteregger's measurements accounts for the lower value of photocurrent observed on this flight. This photocurrent was a constant but negligible factor over much of the altitude range covered (less than 1 percent of the net positive current at the F_2 maximum) but became significant near apogee where it amounted to as much as 25 percent of the measured positive current. The current data presented in figure 7 have not been corrected for photocurrent; however, this correction was made before conversion to ion concentration.

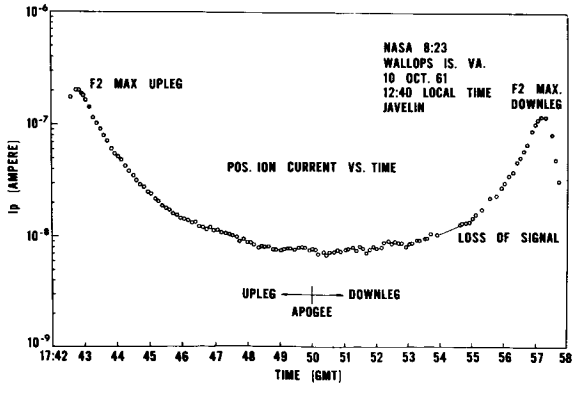


FIGURE 7.—Positive ion current to the electrostatic probe throughout the flight.

Conversion of the currents shown in figure 7 to the related total ion concentration, as discussed in Appendix 2, requires an assumption about the ion species present or an independent determination of it. Below 450 km, O^+ is the dominant F-region ion (Taylor and Brinton, 1961). Above 450 km, however, the ion mass spectrometer data show that He^+ and H^+ become important constituents. The higher mean velocities of these light ions produce a greater positive current to the probe than a like number of oxygen ions and, if not considered, will lead to a total ion concentration that is too high. As a first approximation, however, it is instructive to assume that only O^+ is present and to generate an "upper limit" total concentration, n_{t0}, from the ion current data. The resulting ion concentration profile is shown in figure 8. The ionosonde values of n_e, which are believed most accurate, are shown for comparison. The agreement between n_e and n_{t0} near the F_2 maximum tends to demonstrate the accuracy of equation 12 in Appendix 2.

It is presumed that the part of the profile below 450 km represents a hydrostatic distribution of O^+, the only ion present in significant numbers. Thus, if thermal equilibrium is assumed, the observed scale height of 147 ± 5 km leads to a temperature of $1235°\pm40°$K. An extension of the O^+ distribution to higher altitudes can be carried out rigorously by means of the method outlined by Bauer (1962) and Hanson (1962). For the values of $n(H^+)$ and $n(He^+)$ observed by the spectrometer, however $n(O^+)$ is sufficiently predominant that a linear extrapolation with geometric altitude approximates the hydrostatic distribution. This linear extrapolation of $n(O^+)$ is shown as the dash-line in figure 8. When we compare $n(O^+)$ and n_{t0}, the effect of the helium and hydrogen ions upon the probe current becomes evident just above 500 km and is seen to increase with altitude.

INTERPRETATION OF THE COMBINED RESULTS

The helium and hydrogen ion concentrations can be combined with the oxygen ion concentra-

tion values to permit a check upon the consistency of the two sets of experimental data shown in figures 4 and 8. If we assume that no other ions are present, simple addition of the individual ion concentrations, $n(H^+)$, $n(He^+)$, and $n(O^+)$ provides the total ion concentration n_{t1} shown in figure 9. The same values can be used for

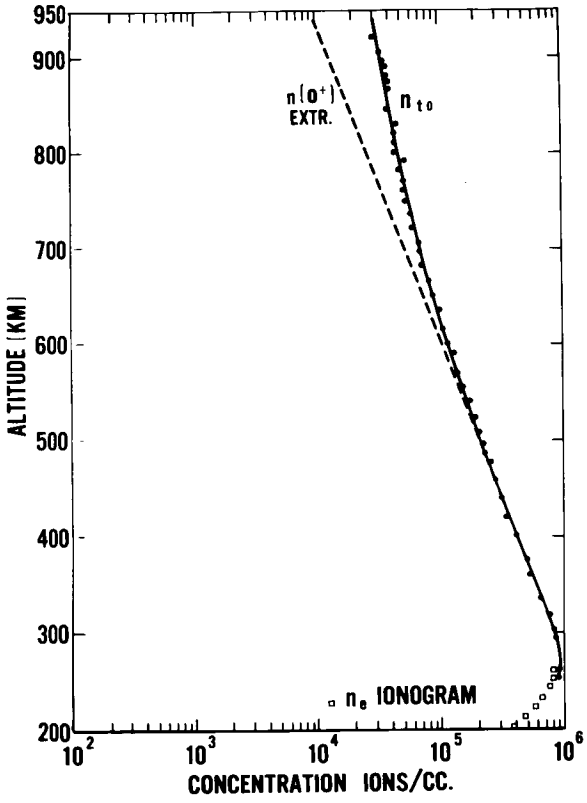


FIGURE 8.—Total ion concentration n_{t0}, derived from the electrostatic probe currents by assuming oxygen ions only, and the oxygen ion concentration $n(O^+)$, derived by extrapolating the n_{t0} measured between 350 and 500 km.

equation 12 to adjust the upper limit total concentration, n_{t0}, for the presence of helium and hydrogen ions. The resulting values of total concentration are labeled n_{t2} in figure 9. Although n_{t1} and n_{t2} are not arrived at entirely independently, the agreement between them demonstrates a satisfying degree of consistency in the flight data from the two experiments. From the electrostatic probe data alone, we would postulate slightly greater concentrations of the light ions than is indicated by $n(H^+)+n(He^+)$;

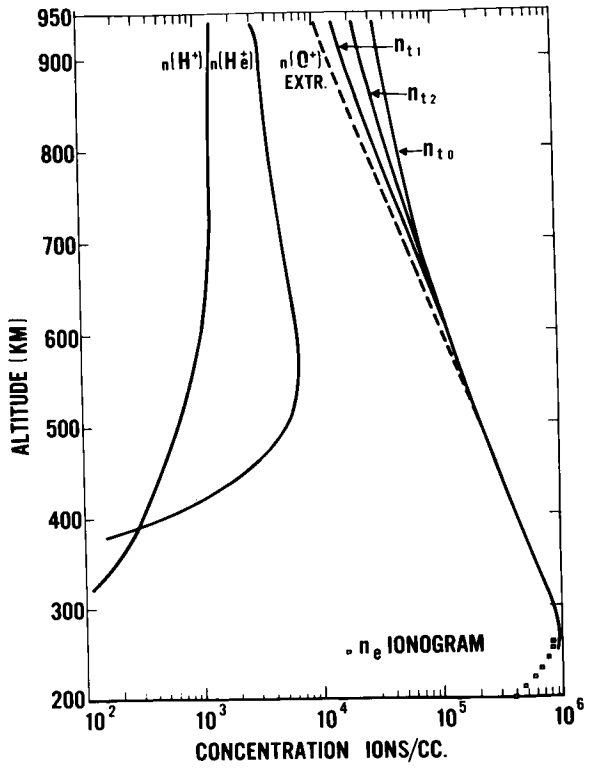


FIGURE 9.—Composite ion concentration profiles including total ion density, n_{t1}, obtained by adding $n(H^+)$, $n(He^+)$, and $n(O^+)$; and n_{t2} derived by correction of n_{t0} for the presence of the lighter ions.

however, the agreement between n_{t1} and n_{t2} is within the combined probable errors of the measurements and the extrapolation method employed in obtaining $n(O^+)$.

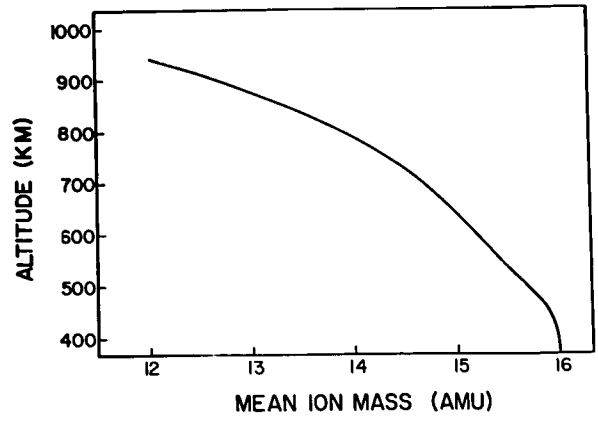


FIGURE 10.—Mean ion mass derived from $n(H^+)$, $n(He^+)$, and $n(O^+)$ values given in figure 9.

The same parameters can be used with (2) to obtain a profile of the mean ion mass m_i that existed in the ionosphere during this flight as shown in figure 10.

$$m_i = (1/n_t) \sum_k n_j m_j \qquad (2)$$

Note that the mean ion mass had decreased from 16 AMU below 400 km to 12 AMU at apogee and was decreasing rapidly there.

DISCUSSION AND CONCLUSIONS

Probably the most significant result of these measurements is the detection of helium and hydrogen ions in the upper F region in substantially the amounts predicted (Nicolet, 1961; Hanson, 1962; Bates and Patterson 1962). However, the occurrence of a helium ion maximum at 575 km is not in accord with ion diffusion theory (Mange, 1960; Bauer, 1962) which predicts that the helium maximum should occur somewhat above the transition level where it becomes the dominant ion. Both the theory and this experiment indicate that the transition level occurs above 1000 km for an ionosphere temperature of 1235°K. We do not interpret the difference between the experimental and theoretical helium distributions as indicating an inadequacy in the diffusion theory, but regard this as evidence of additional geophysical effects that influenced the vertical distribution of ions at the time of this flight. For example, the geomagnetic field and field-alined irregularities, when present, modify the charged particle profiles to various degrees, depending on the latitude and local ionospheric conditions.

The temperature derived from the electrostatic probe data is in agreement with the Harris and Priester (1962) model of gas temperature derived for similar local times and levels of solar acitivity.

The altitude at which He^+ was first detected, and the steep gradient immediately above, are in agreement with Hanson's (1962) predictions based on an ion-atom interchange process between He^+ and N_2. Bates and Patterson have since shown that this process is more likely to involve O_2 than N_2. The comparison of $n(He^+)$ obtained from this flight with the distributions predicted by Bates and Patterson permits us to infer that the rate coefficient for the ion-atom interchange process lies between 3.7×10^{-11} and 2.0×10^{-10} cm^3/sec.

Appendix 1

CONVERSION OF SPECTROMETER COLLECTOR CURRENT TO ION DENSITY

The conversion of the spectrometer collector current to positive ion density involves parameters that are effective inside the sensor as well as those that control the effective collection area associated with the sensor orifice.

A laboratory calibration of the internal sensor parameters is performed to determine the relationship

$$I_c = kI_0 \tag{3}$$

where the spectrometer efficiency $k = \alpha\phi$; α is the combined electrical transparency of the grids determined by measuring the attenuation of an ion beam in passing through a grid system; ϕ is the current efficiency of the analyzer system obtained analytically, and I_0 is the known ion input current measured at the spectrometer orifice. Values of k determined from direct laboratory measurements of I_c and I_0 are in agreement with theoretical predictions of over-all efficiency. This calibration, performed with a controlled beam of positive ions directed along the spectrometer axis, did not include the simulation of effects external to the orifice.

The external effects influencing the collected currents were not subject to direct measurement. To a first-order approximation, the current to the orifice grid can be determined from the expression for a planar probe (Mott-Smith and Langmuir, 1926):

$$I_0 = n_i eaA/2\sqrt{\pi} \tag{4}$$

where n_i is the ion density in ions per cubic centimeter, a is the most probable ion velocity, and A is the area of the orifice grid. In flight, however, the quantity I_0 was influenced by two perturbing factors; first, the negative drawing-in potential E_0 applied to the orifice grid to increase the over-all ion sensitivity; and second, the rocket velocity V_r and angle of attack. These factors combined to modify the input current flow, resulting in the relationship

$$I_0 = (n_i eaA/2\sqrt{\pi}) f(E_0, V_r) \tag{5}$$

Because of the orifice geometry and the resulting fringing field surrounding the orifice grid, we must consider the factors A and $f(E_0, V_r)$ in (5) together to evaluate the resultant effective collection area A_{eff}.

Since the function $f(E_0, V_r)$ was not subject to direct measurement, the value of A_{eff} has been approximated at peak altitude, where the correction for velocity, aspect, and composition is believed to be most reliable. Because of the strong drawing-in field of the first grid (-10 volts) and the relatively low velocity at peak altitude, the term $f(V_r)$ is dropped. An approximation of $f(E_0)$ based on a consideration of the orifice geometry and fringing field results $A_{\text{eff}} = 3A$.

The resulting expression for ion density at peak is

$$n_i = 2\sqrt{\pi} I_c/kea A_{\text{eff}} \tag{6}$$

and by substituting $k = 1.5$ percent, $A_{\text{eff}} = 25.6$ square centimeters, together with appropriate values of I_0 and a, we determined the quantities for $n(\text{H}^+)$ and $n(\text{He}^+)$ at peak.

The energy gained by ambient hydrogen and helium ions because of the relative velocity of the vehicle is of the order of their thermal energies, whereas the energy imparted to the ions by the drawing-in potential is approximately a factor of 100 times the thermal energy. It is assumed, therefore, that effects of vehicle velocity on the collection of light ions, although not constant with altitude, are very small compared to the presumably constant effect of the drawing-in potential. On the basis of these assumptions, then, the ion density distributions in figure 4 have been obtained by applying (6) to the measured ion current distributions in figure 3.

Appendix 2

ELECTROSTATIC PROBE THEORY

If the cylindrical collector were stationary at each point of measurement in the ionosphere, the ion currents recorded throughout the flight would be related in a well-known way to the ion number n_p, mass m_p, charge e, and temperature T_p, as well as the area A and voltage V of the collector. This relationship is expressed for a given ion species by Langmuir's equation for stationary cylindrical probes (Mott-Smith and Langmuir, 1926).

$$i_{ps} = An_p e(kT_p/2\pi m_p)^{1/2} \times (2/\sqrt{\pi})(1+eV/kT_p)^{1/2} \quad (7)$$

Since the term eV/kT_p is normally much greater than 1, equation 7 can be written

$$i_{ps} = An_p e(2eV/\pi^2 m_p)^{1/2} \quad (8)$$

If more than one ion species is present, the current due to each may be considered separately so that the total current is

$$i_{ps} = Ae(2eV/\pi^2)^{1/2} \sum_j n_i/(m_i)^{1/2} \quad (9)$$

For a ternary mixture of hydrogen, helium, and oxygen ions such as that observed on this flight (9) becomes

$$i_{ps} = Ae(2eV/\pi^2 m_{H^+})^{1/2} \times \{n(H^+) + \tfrac{1}{2}[n(He^+)] + \tfrac{1}{4}[n(O^+)]\} \quad (10)$$

In rocket application, however, the probe moves through the plasma at velocities comparable to the mean ion velocities; therefore an additional flux of ions is intercepted or swept out by the collector. In the limit of high velocity, the collector sweeps out ion current proportional to the total number of ions present n_t, their charge e, the translational velocity of the probe W, and the area of the collector projected in the direction of motion $A \sin\theta$. Thus the ion current swept out is given by the equation

$$i_{pw} = n_t eWA \sin\theta \quad (11)$$

where θ is the angle between the probe axis and the velocity vector.

The total ion current to the collector can be viewed as the sum of the random current and the current that is swept out, as given by (12).

$$i_p = i_{ps} + i_{pw} = Ae[(2eV/\pi^2)^{1/2} \times \sum_j n_i/(m_i)^{1/2} + n_t W \sin\theta] \quad (12)$$

This is an oversimplified description of the current collection process, since it assumes that the stationary current (equation 10) and the high-velocity expression (equation 11) apply independently and therefore can be added. Furthermore, the degree of inaccuracy resulting from this assumption is variable, since it depends on both the vehicle velocity and the particular ion species present. For example, the rocket velocity in this flight was always less than the most probable hydrogen ion velocity but exceeded that of the oxygen ions through much of the altitude range. In spite of this, considerable justification for its use, particularly at higher velocities, derives from the agreement evident between the resulting F_2-region ion concentrations and the corresponding electron concentrations derived from the ionosonde records. In this region, i_{ps} and i_{pw} are comparable in magnitude. At higher altitudes, the velocity is less and (12) reduces to (10).

The spinning motion of the rocket permits a simplification of the data analysis. Twice during each rotation, the rocket spin brings the collector perpendicular to the velocity vector W. At these orientations $A \sin\theta$ reduces in effect to A, and the full ion sweeping effect generates maximums in the successive ion current characteristics

that in turn permit this orientation to be easily recognized. The values of ion current measured each time the collector assumed this orientation are shown in figure 7. After correcting for photoemission, we used these currents with (12) to derive the total ion concentration profile given in figure 8.

REFERENCES

BATES, D. R., and PATTERSON, T. N. L., Helium ions in the upper atmosphere, *Planetary Space Sci.*, **9,** 599–605, 1962.

BAUER, S. J., On the structure of the topside ionosphere, *J. Atmospheric Sci.*, **19,** 276–278, 1962.

BOURDEAU, R. E., WHIPPLE, E. C., Jr., DONLEY, J. L., and BAUER, S. J., Experimental evidence for the presence of helium ions based on Explorer 8 satellite data, *J. Geophys. Res.*, **67,** 167–275, 1962.

HANSON, W. B., Upper atmosphere helium ions, *J. Geophys. Res.*, **67,** 183–188, 1962.

HARRIS, I., and PRIESTER, W., Time-dependent structure of the upper atmosphere, *J. Atmospheric Sci.*, **19,** 286–301, 1962.

HINTEREGGER, H. E., DAMON, K. R., and HALL, L. A., Analysis of photoelectrons from solar extreme ultraviolet, *J. Geophys. Res.*, **64,** 961–969, 1959.

JOHNSON, C. Y., Mass determination of ions detected by Bennett ion RF mass spectrometer, *J. Appl. Phys.*, **29,** 740–741, 1958.

MANGE, P., The distribution of minor ions in electrostatic equilibrium in the high atmosphere, *J. Geophys. Res.*, **65,** 3833–3834, 1960.

MOTT-SMITH, H. M., and LANGMUIR, I., The theory of collectors in gaseous discharges, *Phys. Rev.*, **28,** 727–763, 1926.

NICOLET, M., Helium, an important constituent in the lower exosphere, *J. Geophys. Res.*, **66,** 2263–2264, 1961.

SPENCER, N. W., BRACE, L. H., and CARIGNAN, G. R., Electron temperature evidence for nonthermal equilibrium in the ionosphere, *J. Geophys. Res.*, **67,** 157–175, 1962.

TAYLOR, H. A., Jr., and BRINTON, H. C., Atmospheric ion composition measured above Wallops Island, Va., *J. Geophys. Res.*, **66,** 2587–2588, 1961.

———, ———, and SMITH, C. R., Instrumentation for atmospheric composition measurements, *Proc. 8th Ann. Aero-Space Symp.*, pp. 1–14, Instrument Society of America, Washington, 1962.

ELECTRICITY IN THE TERRESTRIAL ATMOSPHERE ABOVE THE EXCHANGE LAYER

ELDEN C. WHIPPLE, JR.

Goddard Space Flight Center

Some of the reactions involving ions and electrons that might occur between the troposphere and the bottom of the ionospheric D-region (20 to 60 km) are discussed. Electrons are produced by cosmic ray ionization and by detachment from negative ions, and are lost by attachment to O_2. Photodetachment predominates during the day, but at night only collisional detachment is effective. Ion–ion recombination accounts for the removal of ions, the Thomson three-body process predominating below 45 km and the two-body neutralization reaction above that altitude. Probable ion and electron densities in this region during quiet solar conditions are presented.

The mechanism of charge collection by bodies in the atmosphere is discussed with respect to two important applications: the effect of dust in providing a recombination surface for ions, and the problem of interpreting current-voltage curves obtained with ion probes. A perturbation solution of the Boltzmann equation to describe ion collection implies two assumptions, each of which becomes questionable at certain altitudes in this region of the atmosphere.

Direct measurements of the electrical properties of this region are difficult to perform from rockets because small currents must be measured accurately yet swiftly and without disturbing the local electrical environment seriously. The data obtained have proved difficult to interpret. Some new techniques are desirable, and possibilities are suggested.

INTRODUCTION

At the Second International Conference on Atmospheric Electricity held at Portsmouth, New Hampshire, in 1958 only one paper was devoted to the electrical properties of the earth's atmosphere above balloon altitudes.[1] Since then, a great amount of data has been obtained through rocket and satellite measurements of the electrical properties of the ionosphere above approximately 90 km. However, there has not been a corresponding advance in knowledge of the region between the altitudes accessible to balloons and the top of the D-region, that is, between about 30 and 90 km. The lack of attention to this part of the atmosphere has been due to several reasons. It naturally has been exciting to explore the properties of the exosphere and interplanetary space, and several scientists who attended the 1958 conference at Portsmouth have changed from studies of the troposphere to the investigation of the upper ionosphere. Another reason, however, for this gap has been the difficulty of making good measurements in this altitude region, which is too low for satellites and too high for balloons.

There have been theoretical advances in this period of time: Nicolet and Aikin,[2] have discussed the mechanism of the D-region formation with respect to the relative importance of ultraviolet, X-ray, and cosmic radiation. Many of the rate coefficients for ion and electron reactions are better known now. Of particular importance here are the attachment and detachment coefficients for reactions of the type

$$e^- + M \rightleftarrows M^-,$$

where M is a neutral atom or molecule.

In the last two to three years there has been an increase in experimental efforts to understand the D-region, in particular, namely that part of the atmosphere below 90 km which can be ionized by solar radiation during quiet solar conditions. This report discusses the probable electrical properties of the atmosphere *below* the D-region, between approximately 20 and 60 km, during normal solar

TABLE 1.—*Rate Coefficients for Reactions 1 Through 6*

Coefficient	Value	References
Electron–ion recombination coefficient α_2	6×10^{-7} cm³/sec (N_2^+) 4×10^{-7} cm³/sec (O_2^+) 3×10^{-8} cm³/sec (NO^+)	6 (*) 7
Mutual neutralization coefficient α_M	1×10^{-8} cm³/sec	3
Photodetachment coefficient d	0.44/sec (O_2^-) 1.4/sec (O^-)	8 2
Collisional detachment coefficient c	4×10^{-20} cm³/sec (with O_2) 1×10^{-14} cm³/sec (with O)	9 4 and 5
Attachment coefficient a	1.5×10^{-30} cm⁶/sec (three-body collision with O_2). 1.3×10^{-15} cm³/sec (radiative attachment to O).	10 11

*Aikin, personal communication (1962).

conditions when cosmic rays are the only source of ionization. Thus, the auroral latitudes where bremsstrahlung from auroral electrons can ionize down to approximately 30 km are not included in the discussion.

ION AND ELECTRON PROCESSES

In addition to the production of electrons by cosmic ray ionization, the following detachment processes from negative ions are potentially important for electron production at these altitudes.

Photodetachment:

$$O_2^- + h\nu \longrightarrow O_2 + e^- , \qquad (1)$$

$$O^- + h\nu \longrightarrow O + e^- \qquad (2)$$

with similar reactions for other negative ions.

Collisional detachment:

$$O_2^- + M \longrightarrow O_2 + M + e^- , \qquad (3)$$

$$O^- + O \longrightarrow O_2 + e^- \qquad (4)$$

and similar reactions for other negative ions.

Free electrons are lost mainly by attachment to O_2, although other reactions undoubtedly do occur:

$$O_2 + O_2 + e^- \longrightarrow O_2^- + O_2 , \qquad (5)$$

$$O + e^- \longrightarrow O^- + h\nu , \qquad (6)$$

$$O_3 + e^- \longrightarrow O_2^- + O . \qquad (7)$$

Positive ions are lost by ion-ion recombination and also by recombination with electrons. The Thomson three-body ion–ion recombination process predominates below about 45 km; but at this altitude the two-body mutual neutralization reaction is equally efficient and, being pressure-independent, is predominant at higher altitudes. The coefficient for the latter reaction is between 10^{-8} and 10^{-7} cm³/sec.[3]

It is advantageous to use the ion equilibrium equation pertinent to clear air in the lower atmosphere to define a "reference" ion density profile denoted as n_0. Thus

$$n_0 = \left(\frac{q}{\alpha_1}\right)^{1/2} , \qquad (8)$$

where q is the cosmic ray ionization rate and α_1 is the effective ion–ion recombination coefficient

$$\alpha_1 = \alpha_T + \alpha_M , \qquad (9)$$

α_T being the Thomson three-body recombination coefficient and α_M the mutual neutralization coefficient. Physically this concept of n_0 is helpful because the presence of free electrons can be considered as a perturbation affecting the ion density profile. In the absence of free electrons, the positive and negative ion densities n_+ and n_- would revert to n_0 in clean air.

Reactions 1 through 7 have been discussed, along with other reactions, by authors interested in D-region ion and electron densities (references 2, 4, and 5, among others). Rate coefficients are available for reactions 1 through 6 and are given in table 1 with their sources.

Figure 1 compares the most probable electron production processes. The negative ion density was assumed to be equal to n_0 for the purpose of computing the curves indicating detachment processes. The values of q used here are those from reference 12, corresponding to 41 degrees north geomagnetic latitude. It is evident that photodetachment predominates during the day if the negative ions are O_2^- or O^-. At night, collisional detachment and cosmic ray ionization both must be taken into account. However, only the collisional detachment process, reaction 3, needs to be considered at night because the atomic oxygen rapidly combines with O_2 to form ozone.[13]

Reaction 5, three-body attachment to molecular oxygen, undoubtedly predominates over the faster atomic oxygen attachment reaction 6 as the most important electron loss process below 60 km because of the low atomic to molecular oxygen ratio. Whitten and Poppoff[5] have concluded that the latter reaction predominates only above 85 km. Electron-ion recombination occurs at a rate that is at least three orders of magnitude smaller than

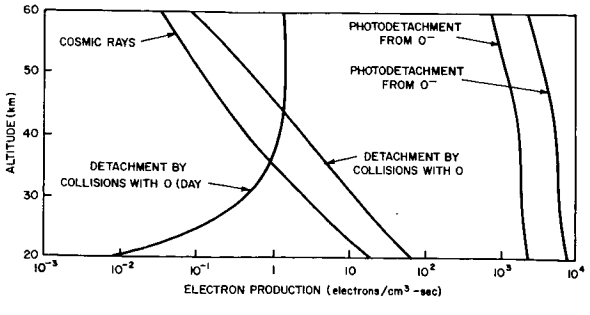


FIGURE 1.—Rate of electron production for various processes.

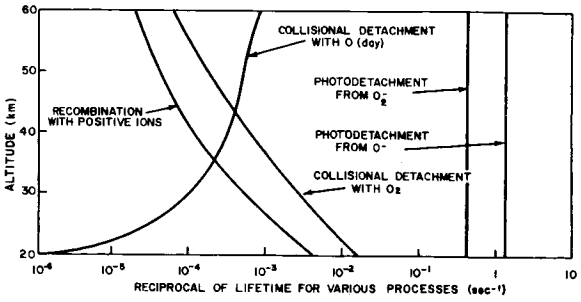


FIGURE 2.—Rate of negative ion loss for various processes, $(1/n_-)\,(dn_-/dt)$.

the rate for attachment to O_2 between 20 and 60 km if a recombination coefficient of 6×10^{-7} cm^3/sec (reference 6) and an ion density of 10^4/cm^3 are assumed—both quite generous assumptions.

The electron density during the day will depend on whether O_2^-, O^-, or some other negative ion is predominant. For O^- to be present in significant quantities, it must be formed by the charge exchange process:

$$O_2^- + O \rightarrow O^- + O_2; \qquad (10)$$

and the rate coefficient must be on the order of 10^{-10} cm^3/sec. This coefficient, although not known, may be as large as this, in which case O^- may be important during the day. However, this disagrees with the sunrise–sunset effect in the D-region.[5]

Loss rates for negative ions are compared in figure 2, where the reciprocal of the ion lifetime for various processes is plotted against altitude. It is clear that photodetachment is the only process that needs to be considered in the day if the ion is O_2^- or O^-. At night, both collisional detachment and recombination with positive ions must be taken into account although the former is not an efficient loss mechanism because of the rapid reattachment of the electrons to O_2. There is some question as to how much the detachment coefficients of table 1 change when the reduced solar spectrum at lower altitudes is used, rather than that above the atmosphere. The cross sections for detachment from O_2^- and O^- are almost entirely in the visible and infrared, implying little effect of the ultraviolet absorption by ozone on the coefficients. This again disagrees with the "sunrise" effect, and Whitten and Poppoff,[5] have

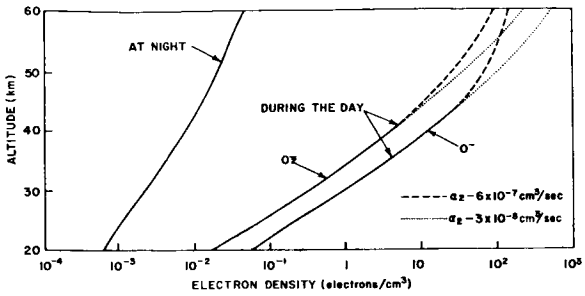


FIGURE 3.—Electron densities resulting from various choices for positive and negative ion species.

speculated that the O_2^- ion may be in a lower state in the atmosphere than during the experimental determination of the coefficient so that radiation between 2500 and 3000 A is responsible for detachment in the atmosphere.

It is felt that the most likely negative ion is O_2^-. An upper limit on the electron density can be obtained by assuming O^- instead of O_2^-. The values of electron density obtained for these negative ion assumptions are compared in figure 3, along with a maximum nighttime value. Other species of negative ions (O_3^-, NO_2^-, etc.) may occur, but more information on the rate coefficients involved in their production is needed before anything definite can be said about their importance. These have lower photodetachment coefficients and should yield lower daytime electron density values. Electron densities were computed for positive ion choices of N_2^+ and NO^+ so that the effect of different electron–ion recombination coefficients could be seen. Intermediate values of electron density would be obtained if the positive ion were O_2^+.

Positive ion densities, computed for the same choices of positive and negative ion species, are presented in figure 4. The following equilibrium equations were used:

for positive ions,

$$q = \alpha_1 n_+ n_- + \alpha_2 n_+ n_e \quad \text{(day or night)}; \quad (11)$$

for electrons (N being the number of O_2 molecules/cm^3),

$$q + dn_- = aN^2 n_e + \alpha_2 n_+ n_e \quad \text{(day)}, \quad (12)$$

$$q + cn_- N = aN^2 n_e + \alpha_2 n_+ n_e \quad \text{(night)}; \quad (13)$$

for negative ions,

$$aN^2 n_e = dn_- + \alpha_1 n_+ n_- \quad \text{(day)}, \quad (14)$$

$$aN^2 n_e = cn_- N + \alpha_1 n_+ n_- \quad \text{(night)}; \quad (15)$$

and the neutrality equation,

$$n_+ = n_e + n_- . \quad (16)$$

Solutions obtained during the day are

$$\left.\begin{aligned} n_+ &\approx \sqrt{\frac{q}{\alpha_1}} \sqrt{1 + \frac{d(\alpha_1 - \alpha_2)}{aN^2 \alpha_1 + d\alpha_2}} \\ n_e &\approx \frac{q + dn_+}{aN^2 + d} \end{aligned}\right\} \quad (17)$$

and those obtained at night are

$$\left.\begin{aligned} n_+ &\approx \sqrt{\frac{q}{\alpha_1}} \sqrt{\frac{aN^2 \alpha_1 + cN\alpha_1}{aN^2 \alpha_1 + cN\alpha_2}} \rightarrow \sqrt{\frac{q}{\alpha_1}} \\ n_e &\approx \frac{q + cNn_+}{aN^2 + cN} \rightarrow \frac{q + cNn_+}{aN^2} \end{aligned}\right\} \quad (18)$$

During the day the electron density is essentially proportional to the appropriate negative ion photodetachment coefficient. Above 45 km it is mildly sensitive to the nature of the positive ion. If the negative ion were not O_2^- or O^- but something with a low photodetachment coefficient, then the electron density would approach the negligible nighttime values.

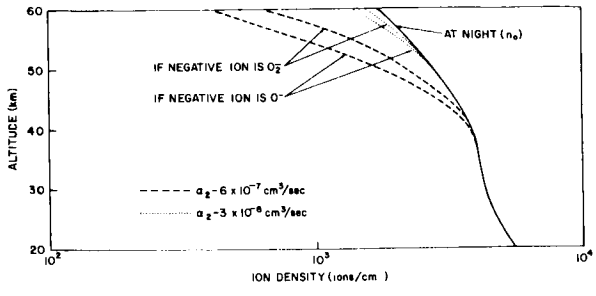


FIGURE 4.—Positive ion densities resulting from various choices for positive and negative ion species.

The positive ion density is sensitive to the nature of the ion only at altitudes where recombination with electrons can compete with ion–ion recombination. Hence the density profile corresponding to the choice of N_2^+ and O^- represents a sort of lower limit to the ion density at this latitude unless the true electron–ion recombination coefficient is even higher than 6×10^{-7} cm^3/sec.

Nicolet and Aikin[2] have obtained electron and ion densities for the D-region above 60 km under quiet solar conditions. Their values at 60 km fall within the range of values obtained here for the same altitude.

In the lower atmosphere, ions form clusters with neutral molecules so that the effective ion mass is larger than that of the simple molecular ion. This tendency to cluster would lower the value of the effective recombination coefficient. The altitude at which clustering may become significant is not known, but it could perhaps be detected by measuring the ion density at night. An abrupt decrease with height in the ion density at a certain altitude might be indicative of the increase in the value of the recombination coefficient corresponding to cluster disassociation.

These considerations all have been for a clean atmosphere where the diffusion of electrons and ions to dust is negligible. Before considering this effect on the electron and ion densities, the mechanism of charge collection by bodies in this part of the atmosphere will be discussed.

CHARGE COLLECTION BY BODIES IN THE ATMOSPHERE

The theory of charge collection by bodies in the atmosphere has two important applications: One is the role that dust plays in the ion equilibrium in the atmosphere by providing a recombination surface; the other is in the use of probes where a current-voltage characteristic is interpreted in terms of the atmospheric electrical properties.

The atmosphere up through the D-region can be divided into two distinct regions as far as this problem is concerned: (1) lower altitudes where the ionic mean free path L is much smaller than the body dimension, and (2) higher altitudes where this is no longer true. The two cases where L is either much larger or much smaller than the body dimension have been treated extensively in the literature; but this is not true for the transition region, where L is the same order of magnitude as the collecting body. The transition region occurs at different altitudes for the two applications of concern here because of the difference in size between dust particles and probes. A typical probe is much larger than the mean free path to about 70 km, where L is about 0.1 cm. On the other hand, the transition altitude for a 10-micron dust particle occurs at 35 km.

Combination coefficients for the diffusion of ions to dust in the lower atmosphere have been obtained by Gunn[14] and Bricard,[15] and a personal communication from Bricard in 1963. Use of probes in the lower atmosphere is also well understood (references 16, 17, 18, and 19, among others). All of these applications have been treated in essentially the same way—by considering the current density $\vec{J}$ to consist of separate terms corresponding to conduction, diffusion, and perhaps convection. The total current to the body in question is obtained by expressing $\vec{J}$ in terms of the local field quantities (electric field, particle concentration and its gradient, airflow, etc) and integrating over the appropriate collecting area. Space charge effects can be taken into account through Poisson's equation.

The use of the current density in terms of transport coefficients is given by

$$\vec{J}=\sigma\vec{E}-eD\vec{\nabla}n+ne\vec{W},$$

where σ is the conductivity, $\vec{E}$ the electric field, e the unit charge, D the diffusion coefficient, and $\vec{W}$ a convection velocity. The use of equation 19 implies at least two assumptions, both of which can become questionable at certain altitudes in the atmosphere. This can be made clear by considering how this description of $\vec{J}$ is obtained. Boltzmann's equation in the steady state for the particle velocity distribution $f(\vec{v})$ is solved by assuming a solution of the form $f=f_0+f_1$ and replacing the collision term by either the proper collision integral or the approximation $-(f-f_0)/\tau$, where f_0 is the normal Maxwell distribution and τ is an appropriate relaxation time. The quantity f_1 is assumed to be small so that its derivatives and powers can be neglected. Thus, for example, in a one-dimensional case where an electric field

is the only external force we obtain

$$f_1 = f_0 - \tau\left(v_x \frac{\partial f_0}{\partial x} + \frac{eE}{m}\frac{\partial f_0}{\partial v_x}\right) \qquad (20)$$

where $\partial f_1/\partial x$ and $\partial f_1/\partial v_x$ have been neglected. The current density $\vec{J}$ is then obtained by integrating over the velocity distribution:

$$\vec{J} = e\int \vec{v}\, f(\vec{v})\, d\vec{v} \qquad (21)$$

If f_0 is assumed proportional to

$$n(x)\, \exp[-v_x^2/(2kT/m)],$$

the conditions

$$\frac{\partial f_1}{\partial x} \ll \frac{\partial f_0}{\partial x} \text{ and } \frac{\partial f_1}{\partial v_x} \ll \frac{\partial f_0}{\partial v_x} \qquad (22)$$

become

$$eEL \ll kT \text{ and } \frac{1}{n}\frac{\partial n}{\partial x} \ll \frac{1}{L} \qquad (23)$$

That is, (1) the energy gained in one mean free path by an ion in an electric field must be less than its kinetic energy; and (2) the ion density must not change appreciably over one mean free path.

Consider now a typical probe such as a Gerdien condenser designed for a rocket-borne experiment. Let the maximum electric field in the condenser be about 5 volts/cm. For the first condition to be satisfied.

$$L \ll kT/eE \approx \frac{0.025 \text{ volt}}{5 \text{ volts/cm}} = 0.005 \text{ cm}.$$

The mean free path is less than this only below 45 km.

The second condition is not violated as easily in the lower atmosphere. However, during events such as solar flares there will be abnormal ionization in and below the D-region which could be large enough so that space charge sheaths may be formed around probes. In such a case, the ion density will change significantly over a Debye length. For example, during a strong solar flare the ion density at 75 km could be as high as $2\times10^4/\text{cm}^3$ (reference 2). The corresponding Debye length is about 1 cm, which is only five times the mean free path at that altitude. Consequently, care should be used in interpreting probe data under such conditions.

The current to a charged sphere at rest in that part of the atmosphere where the mean free path is very large compared with the particle radius was first obtained by Mott-Smith and Langmuir.[20] Their solution for charges that are attracted is appropriate to the case where the particle radius is much smaller than the Debye length—the so-called orbital-motion-limited current case.[21] The currents are

$$I_\pm = \pm\pi r^2 e C_\pm N_\pm [1 \mp (\Phi e/kT)]$$

(attracted charges), (24)

and

$$I_\pm = \pm\pi r^2 e C_\pm N_\pm \exp(\mp\phi e/kT)$$

(repelled charges), (25)

where r is the particle radius, $C_\pm$ the mean thermal velocity of the charged particle, and Φ the potential of the particle with respect to the atmosphere. The corresponding combination coefficients are

$$\beta_1 = \pi r^2 C_\pm [1 \mp (\Phi e/kT)]$$

(attracted charges), (26)

$$\beta_2 = \pi r^2 C_\pm \exp(\mp\phi e/kT)$$

(repelled charges). (27)

The combination coefficients obtained in the lower atmosphere by Gunn,[14] and Bricard (private communication), are expressed as functions of the mean free path, and we would expect that in the limit, as the mean free path becomes very large, these would approach the expressions appropriate in the upper atmosphere (equations 26 and 27). However, this is true only for the case when the particles have zero charge on them. Examination of the derivation of the coefficients for the lower atmosphere shows that the boundary condition used at the particle surface can be improved. For example, Gunn,[14] equates the current collected by the particle to the effusion current at a surface situated one mean free path from the particle surface. These effusion currents are identical to the currents given in equations 24 and 25 for the case of zero charge. When we use the full expressions, equations 24 and 25, for the boundary conditions and replace Φ by $\Delta\Phi$, the

potential drop between the particle and the imaginary surface at one mean free path, we obtain the following combination coefficients that have the correct limit as $L \to \infty$ and can be used in the the transition region in the atmosphere where $L \approx r$:

$$\beta_1 = \frac{\pi r^2 C_\pm \left(1 \mp \frac{Qe}{kTr} \pm \frac{Qe}{kT(r+\Delta)}\right) e^{\mp \frac{Qe}{kT(r+\Delta)}}}{1 \pm \frac{r^2 C_\pm}{4QU_\pm}\left[1 - e^{\mp \frac{eQ}{kT(r+\Delta)}}\right]} \quad (28)$$

and

$$\beta_2 = \frac{\pi r^2 C_\pm e^{\mp \frac{Qe}{kTr}}}{1 \pm \frac{r^2 C_\pm}{4QU_\pm}\left[1 - e^{\mp \frac{eQ}{kT(r+\Delta)}}\right]} \quad (29)$$

where $Q = \Phi r$ is the charge on the particle of radius r, U is the ion mobility, and Δ is the corrected mean free path.[22] The coefficients β_0, β_1,

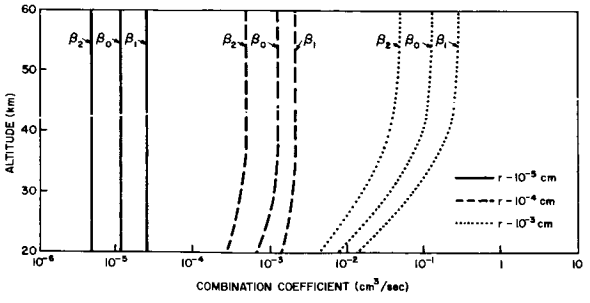


FIGURE 5.—Combination coefficients for loss of ions to dust.

and β_2, where β_0 is the combination coefficient for zero charge on a particle, are plotted against altitude in figure 5. The charge Q has been chosen so that $(\Phi e/kT) = \pm 1$ for β_1 and β_2, respectively.

The effect of dust in reducing the daytime ion density is shown in table 2, where values of ion and electron density at 60 km in clean air are compared with the values obtained when a dust density of 1 particle/cm³ is assumed. To obtain these values, equations 11, 12, and 14 were modified by adding a term corresponding to the diffusion to dust process. An additional equation

TABLE 2.—*Effects of One Dust Particle/cm³ on Ion and Electron Densities**

	No dust	$r = 10^{-5}$ cm	$r = 10^{-4}$ cm	$r = 10^{-3}$ cm
n_-	625	466	15	0.271
n_+	544	403	0.8	8×10^{-3}
n_e	81	60	0.1	1×10^{-3}
Q/e	---------	−3	−14.1	−0.262

* α_2 has been assumed to be 6×10^{-7} cm³ sec.

was obtained by setting the net current to a dust particle equal to zero. Finally, equation 16 had a term added corresponding to the charge per cm³ residing on dust particles. The effect of photoemission from the dust particles in sunlight in increasing the electron production rate has been neglected.

The problem of measuring conductivity with ion probes on a rocket has been discussed by Bourdeau, Whipple, and Clark,[12] and by Smith (personal communication, 1963). Smith gives expressions for the current to a probe flush with the rocket skin, taking into account the fact that the measurement actually involves a double probe since the rocket body, as well as the probe itself, collects current. It is not clear how this probe behaves in the transition region where the mean free path is large.

As an example of the kind of problem encountered in trying to predict the current–voltage characteristics of a probe in this region of the atmosphere and also as an indication of how the problem might be attacked, consider a hypothetical probe that combines the features of a Gerdien condenser and an ion trap, as shown in figure 6. The probe consists of a hollow cylinder like a

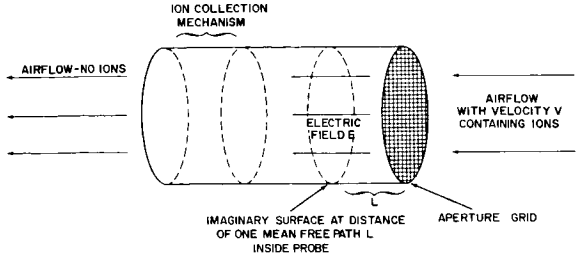


FIGURE 6.—Hypothetical probe for collecting ions.

Gerdien condenser mounted so that there is an airflow of known velocity V entering the aperture as shown. A wire grid through which the air must flow is mounted at the aperture and is at the same potential as the outside of the probe, which we will assume to be the potential of the surrounding atmosphere. Inside the probe there is some means of collecting ions of the desired polarity so that none leave the cylinder with the air escaping from the other end. The ion collection mechanism is such that inside the cylinder there is a uniform attractive electric field at the aperture grid normal to the plane of the grid.

The measured ion current corresponds to what is usually called the *saturation current*; that is, it is determined by the rate at which ions enter the cylinder. In the lower atmosphere when the electric field is small, this current is given by

$$I = neVA, \tag{30}$$

where A is the cross-sectional area of this aperture.

Now it is well known that an ion trap mounted on the forward face of a satellite measures an ion current that is given by the identical expression.[23] In spite of the vast difference in the mean free path in these two applications the current is of the same form because the same physical requirement is met, namely, that the net distance traveled by the ion in a given time with respect to the medium is very small compared with the distance the ion is carried along by the medium in the same time. In the satellite case this is true because the satellite velocity is much greater than the random thermal velocity of the ion. In the lower atmosphere this is true because the short mean free path restricts the net distance that the ion travels even though the thermal velocity may be much greater than the airflow velocity.

The fact that equation 30 is obtained in the satellite case by integrating equation 21 suggests that it should be possible to obtain it in the same way for the other case, and hence for intermediate cases as well. This is indeed true if we are careful to take into account the ions that penetrate the grid once but then suffer a collision so that thay are deflected back through the grid to escape. The following treatment indicates in a heuristic way how this might be done.

We assume that the ion velocity distribution at the aperature is Maxwellian except for the stream velocity. Thus,

$$f_1(v_x) = \frac{n}{2a\sqrt{\pi}} e^{-(v_x+V)^2/a^2} \tag{31}$$

where a is the most probable ion thermal velocity. After the ions enter the grid, they are accelerated by the electric field; but because of collisions the net result is that the ions acquire a drift velocity w superimposed on the thermal and stream velocities. The resulting distribution is given by

$$f_2(v_x) = \frac{n}{2a\sqrt{\pi}} e^{-(v_x+V+w)^2/a^2} \tag{32}$$

We now further assume that equation 32 is descriptive of the distribution after the ions have suffered only one collision. In particular, $f_2(v_x)$ is taken to describe the distribution at a distance of one mean free path inside the cylinder from the aperture grid.

The initial current I_1 that enters the cylinder is given by equation 21 integrated over $f_1(v_x)$:

$$I_1 = neVA\left[\frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\,\frac{V}{a} + \frac{a}{2V\sqrt{\pi}} e^{-(V/a)^2}\right] \tag{33}$$

The return current I_2, consisting of those ions that escape, is computed by determining the reverse current at the imaginary surface situated one mean free path inside the cylinder, taking into account the fact that not all the ions will have sufficient energy to overcome the potential rise from this surface to the grid. Thus,

$$I_2 = Ae \int_{-\infty}^{-v_2} v_x f_2(v_x)\, dv_x \tag{34}$$

where v_2 is given by

$$v_2 = \sqrt{\frac{2eEL}{m}}\,. \tag{35}$$

The result is

$$I_2 = -neA(V+w)\left[\frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\,x - \frac{a}{2\sqrt{\pi}(V+w)} e^{-x^2}\right] \tag{36}$$

where

$$x = -\frac{1}{a}(V+w+v_2)\,. \tag{37}$$

Hence, the desired current is given by the difference

$$I = I_1 - I_2. \tag{38}$$

This current normalized to the value given by equation 30 is plotted against the mean free path parameter $2EeL/ma^2$ in figure 7 for various values of V/a. The drift velocity w has been taken from Wannier's[24] equations with the assumption that the ion mass is equal to that of the neutral molecule. The corresponding altitude for a typical field of 1 volt/cm and a temperature of 300°K is also indicated. It is apparent that, as long as the stream velocity—i.e., the rocket or satellite velocity—is greater than the most probable ion velocity, then the collected current is independent of the altitude and is given by equation 30. This is true for ion collection but not for electron collection from both rockets and satellites. For electrons, the deviation from equation 30 is significant above 80 km for a rocket moving at 1 km/sec. It should be noted that, in addition to the neglect of any expected vehicle potential, this treatment neglects the entrance of particles at the exit end of the cylinder against the airstream.

EXPERIMENTAL PROBLEMS

It has not yet been possible to test experimentally the ideas presented in the first section as to the ion and electron densities in this part of the atmosphere above balloon altitudes. The results of only two measurements have been published so far (reference 12 and a personal communication from Smith, 1963), and the interpretation of these measurements can be questioned. It is felt that the greatest need at present is the development of reliable measuring techniques. This means either that the same quantity should be measured simultaneously in different ways or that enough simultaneous measurements of different quantities should be made so that some independent requirement such as the neutrality equation can be used to verify the results.

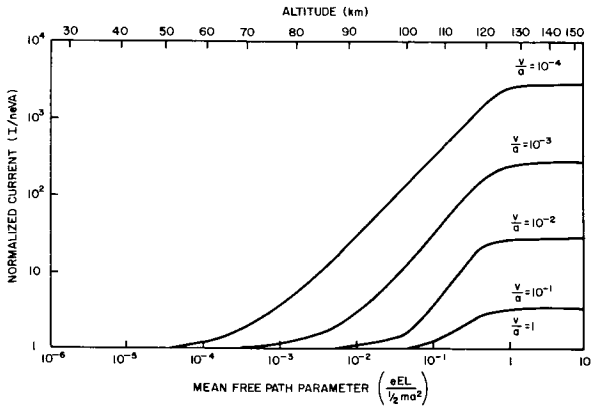


FIGURE 7.—Ion current collected by the probe of figure 6.

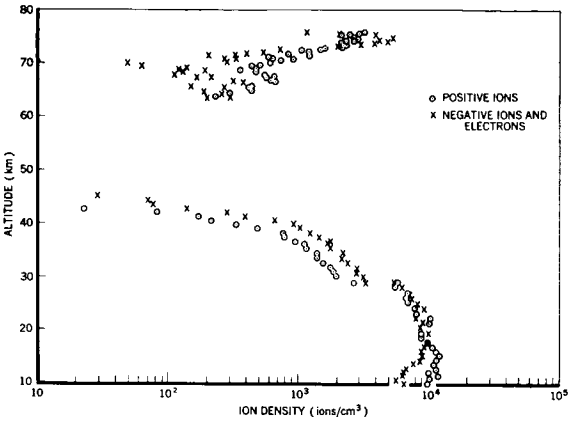


FIGURE 8.—Charged particle densities at Fort Churchill, Canada. NASA rocket 1-02, 1150 CST, November 27, 1960 (ascent).

To illustrate further some of the difficulties of data interpretation, some previously unpublished results are presented in figures 8 and 9. Figure 8 shows positive and negative ion densities measured with two independent Gerdien condensers on an Aerobee rocket in Fort Churchill, Canada. The condensers had a constant potential of approximately ±100 volts applied between the electrodes. The corresponding critical mobility for the condensers was computed to be equal to the actual small-ion mobility at 29 km (94.5 cm²/volt sec). Below this altitude, the measured currents that increased with altitude were assumed to be proportional to the ionic conductivity, and the ion density shown was computed by assuming a mobility corresponding to small ions. Above 29 km the measured current decreased, and the ion densities were computed by assuming the current to be proportional to the ion density and the volume airflow through the condenser. There is a discrepancy of about a factor of 2 in the ion densities obtained by the two methods at the

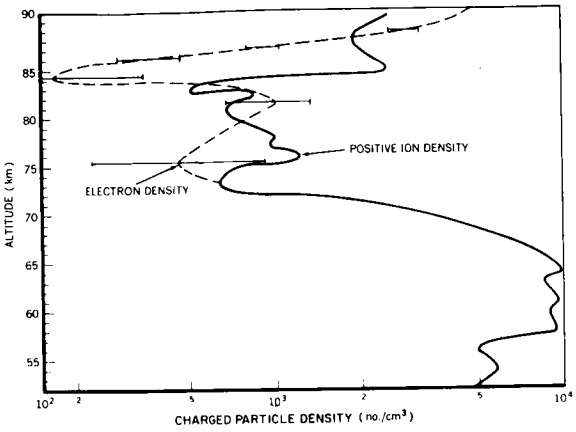


FIGURE 9.—Charged particle densities at Wallops Island, Va. NASA rocket 10.74, 1630 EST, December 21, 1961 (ascent).

critical altitude of 29 km. This may be due to the shock wave in front of the rocket, which would reduce the airflow through the condenser. However, this effect does not appear adequate to explain the continued decrease in current with altitude, since the rocket is decelerating in this region. At first glance it appears unreasonable that the ion density should decrease to such low values. However, the instrumentation performed properly, since the data obtained during the rocket descent reproduced that obtained on the ascent. The same phenomenon was also observed four days earlier on a similar rocket fired at night. A dust density of about one 1-micron particle/cm³ is adequate above 30 km to account for the observed decrease in ion density. On the other hand, this effect may be due to uncertainties in how to interpret the behavior of the Gerdien condenser.

In contrast to this, the measurements between 23 and 29 km indicate relatively clean air. The negative-to-positive-ion conductivity ratio, which fluctuated irregularly from 0.61 to 1.37 between 15 and 22 km with an average of 1.07, increased noticeably at 23 km and averaged 1.40 (approximately the small-ion mobility ratio) between 23 and 29 km.

Figure 9 shows positive ion and electron densities obtained from a Nike-Cajun rocket at Wallops Island, Virginia, in December 1961. The ion densities were obtained by the author from the measured conductivity using the probe described by Smith (personal communication, 1963), and are quite sensitive to the assumed ion mobility. Here Dalgarno's[25] estimate of mobility for N_2^+ in air was used. Electron densities were computed from measurements of the differential absorption of the ordinary and extraordinary propagation modes of a 3-Mc signal transmitted from the ground to the rocket (J. Troim, personal communication, 1962). The altitude-dependence of the electron collision frequency that is necessary for the computation is that given by Kane.[26] There is only general agreement within an order of magnitude between the ion and electron results between 72 and 83 km—which is all that should be expected, considering the error bars on the electron density measurement and the uncertainty in the computation of the positive ion density. The detailed shapes of the profiles do not correlate well, nor should they if dust particles or negative ions were present in significant quantities.

Measurements in this part of the atmosphere would be greatly simplified if a slowly moving observation platform were to be developed. What is needed is a drag device such that a payload ejected above 80 km would reach a terminal velocity of less than Mach 1 at 70 km. In addition to eliminating undesirable shock wave complications from experiments, such a platform would enable measurements to be made over a much longer period of time. Instrumentation response times would not have to be so short and, consequently, smaller currents could be measured. At lower altitudes the present balloon capabilities of reaching 45 km should be exploited.

Other kinds of experiments that are valid above or below this part of the atmosphere should be extended. Ion mass spectrometers are useful now only above about 90 km. A cyclotron resonance ion spectrometer that uses a superconducting magnet is being developed at present and may be useful down as low as 60 km (J. A. Kane, personal communication, 1963). The small atmospheric ions in the troposphere have never been positively identified. A simultaneous determination of the ion mass spectrum and mobility spectrum at, say, 60 km could lead by inference to ion identification at lower altitudes.

Methods now in use for determining dust con-

centrations and radii should be extended above balloon altitudes to establish whether dust has a significant role in the ion equilibrium problem. Electric field measurement techniques also should be extended to higher altitudes. It will be necessary to have a sensitivity such that fields considerably less than 1 volt/meter can be measured. It would be of great interest, for instance, to measure the electric field as a function of altitude between the top of a thunderstorm and the ionosphere. Such an experiment would verify conductivity measurements and would indicate the current density pattern if the field direction could be established as well.

Finally, there is a need for good cosmic ray ionization measurements above balloon altitudes. Only estimates of this quantity, based on counter results, are available at present. It will be a problem to take into account the effect of secondaries produced in structures near the experiment.

In addition to the measurements just listed, there is a need for more laboratory measurements of rate coefficients, particularly for charge exchange reactions and ion processes involving ozone.

CONCLUSIONS

It is emphasized again that the primary need at present is for reliable measurement techniques in this altitude region. Techniques involving ion or electron collection should be examined carefully before the results are interpreted in terms of geophysical quantities. Once such techniques have been developed, they should be used to establish first the normal electrical characteristics of this region during quiet conditions as defined in the introduction. The predominant processes should be established and the ion species identified. Only when this has been done will it be really possible to investigate and understand abnormal conditions that occur during solar flares or at auroral latitudes.

ACKNOWLEDGMENTS

The author would like to thank A. C. Aikin and R. E. Bourdeau for their encouragement and helpful discussions during the preparation of this paper.

REFERENCES

1. Hill, E. L., "Free Electrons in the Lower Atmosphere," in: *Recent Advances in Atmospheric Electricity*, ed. by L. G. Smith, New York: Pergamon Press, 1958, pp. 469–473.
2. Nicolet, M., and Aikin, A. C., "The Formation of the D Region of the Ionosphere," *J. Geophys. Res.*, **65**(5): 1469–1483, May 1960.
3. Nawrocki, P. J., and Papa, R., "Atmospheric Processes," Geophysics Corp. of America, Bedford, Mass., Report No. 61-37-A, 1961.
4. Whitten, R. C., and Poppoff, I. G., "Associative Detachment in the D Region," *J. Geophys. Res.*, **67**(3): 1183–1185, March 1962.
5. Whitten, R. C., and Poppoff, I. G., "Determination of Ionospheric Recombination Coefficients," Stanford Res. Inst., Menlo Park, Calif., Semiannual Reports No. 2 and 3, 1962.
6. Kasner, W. H., Rogers, W. A., and Biondi, M. A., "Electron-Ion Recombination Coefficients in Nitrogen and in Oxygen," *Phys. Rev. Letters*, **7**(8): 321–323, October 15, 1961.
7. Smith, L. G., "Rocket Measurements of Electron Density and Temperature in the Nighttime Ionosphere," Geophysics Corp. of America, Bedford, Mass., GCA TR-62-1-N, January 1962.
8. Bailey, D. K., "Abnormal Ionization in the Lower Ionosphere Associated with Cosmic-Ray Flux Enhancements," *Proc. I.R.E.*, **47**(2): 255–266, February 1959.
9. Phelps, A. V., and Pack, J. L., "Collisional Detachment in Molecular Oxygen," *Phys. Rev. Letters*, **6**(3): 111–113, February 1, 1961.
10. Chanin, L. M., Phelps, A. V., and Biondi, M. A., "Measurement of the Attachment of Slow Electrons in Oxygen," *Phys. Rev. Letters*, **2**(8): 344–346, April 15, 1959.
11. Branscomb, L. M., and Burch, D. S., et al., "Photodetachment Cross Section and the Electron Affinity of Atomic Oxygen," *Phys. Rev.*, **111**(2): 504–513, July 15, 1958.
12. Bourdeau, R. E., Whipple, E. C., Jr., and Clark, J. F., "Analytic and Experimental Electrical Conductivity Between the Stratosphere and the Ionosphere," *J. Geophys. Res.*, **64**(10): 1363–1370, October 1959.
13. Barth, C. A., "Nitrogen and Oxygen Atomic Reactions in the Chemosphere," in: *Chemical Reactions in the Lower and Upper Atmosphere* (Stanford Research Institute), New York: Interscience Publishers, 1961, pp. 303–326.
14. Gunn, R., "Diffusion Charging of Atmospheric Droplets by Ions, and the Resulting Combination Coefficients," *J. Meteorol.*, **11**(5): 339–347, October 1954.
15. Bricard, J., "Ionic Equilibrium in the Lower Atmosphere," *J. Geophys. Res.*, **54**, 39–52, March 1949 (in French).

16. GERDIEN, H., "Absolute Determination of Electric Conductivity and Specific Velocity of the Ions in the Atmosphere," *Phys. Zeitschr.*, **4,** 632–635, September 1, 1903.
17. SWANN, W. F. G., "The Theory of Electrical Dispersion into the Free Atmosphere, with a Discussion of the Theory of the Gerdien Conductivity Apparatus, and of the Theory of the Collection of Radioactive Deposit by a Charged Conductor," *Terrest. Magnetism and Atmos. Elect.*, **19(2)**: 81–92, June 1914.
18. ISRAEL, H., "Theory and Method of Determination of Size of Ions in the Atmosphere," *Gerlands Beitr. Z. Geophys.*, **31**(1–3): 173–216, 1931.
19. KASEMIR, H. W., and RUHNKE, L. H., "Antenna Problems of Measurements of the Air-Earth Current," in: *Recent Advances in Atmospheric Electricity*, ed. by L. G. Smith, New York: Pergamon Press, 1958, pp. 137–147.
20. MOTT-SMITH, H. M., and LANGMUIR, I., "Theory of Collectors in Gaseous Discharges," *Phys. Rev.*, **28,** 727–763, October 1926.
21. KANAL, M., "Theory of Current Collection of Moving Spherical Probes," Space Physics Res. Lab., U. of Michigan, Ann Arbor, Sci. Rept. JS–5, April 1962.
22. BRICARD, J., "La Fixation des Petits Ions Atmospheriques sur les Aérosols Ultra-Fins," *Geofisica Pura e Applicata*, **51,** 237–242, January/April 1962.
23. BOURDEAU, R. E., WHIPPLE, E. C., Jr., et al., "Measurements of Sheath Currents and Equilibrium Potential on the Explorer VIII Satellite," *J. Astronaut. Sci.*, **8**(3): 65–73, 1961.
24. WANNIER, G. H., "Motion of Gaseous Ions in Strong Electric Fields," *Bell Syst. Tech. J.*, **32,** 170–254, January 1953.
25. DALGARNO, A., "Charged Particles in the Upper Atmosphere," *Annales de Geophys.*, **17**(1): 16–49, January/March 1961.
26. KANE, J. A., "Reevaluation of Ionospheric Electron Densities and Collision Frequencies Derived from Rocket Measurements of Refractive Index and Attenuation," NASA Technical Note D–503, November 1960.

V. FIELDS AND PARTICLES

THE MAGNETIC FIELD OF THE QUIET-TIME PROTON BELT

SYUN-ICHI AKASOFU

University of Alaska

JOSEPH C. CAIN

Goddard Space Flight Center

SYDNEY CHAPMAN*

University of Alaska

The distortion of the earth's magnetic field produced by the proton belt is discussed. The magnetic field is calculated numerically, to a first approximation, for an analogous model belt, in a steady state. In the equatorial plane, at the earth's surface, it is estimated that the magnetic field produced by this belt is of order 38 gammas; it is directed southward. The maximum field reduction is of the order of 72 gammas at 4.1 earth radii; this is 15.5% of the dipole field intensity at this point. Beyond 6.7 earth radii, the belt increases the earth's field.

INTRODUCTION

Recent improved measurements by satellites, especially Explorer XII (1961υ), have revealed an extensive belt of low energy protons (150 Kev–4.5 Mev), encircling the earth.[1] A typical flux distribution along one of the paths of Explorer XII is shown in figure 1a for magnetically quiet conditions as a function of radial distance from the earth's center: the path considered lay close to the equatorial plane beyond $3.5a$ (where a is the radius of the earth). In this paper, we calculate the magnetic field produced by a model belt which is similar to the proton belt.

THE MODEL PROTON BELT

When conditions in the belt are steady, the magnetic field produced is determined by the number density n and the velocity distribution $f(v)$ of the particles of the belt, as a function of r_e, the distance from the earth's center along an equatorial radius.[2,3] The velocity distribution involves both the energy spectrum for each kind of particle (protons and electrons), and the pitch-angle distribution; but in the present calculations we shall consider only protons with a particular energy E (or speed v).

The number density distribution is given by:

$$n = n_0 \exp(-g_1^2 z^2)$$

($z<0$ for the inner part of the belt); (1)

$$n = n_0 \exp(-g_2^2 z^2)$$

($z>0$ for the outer part of the belt), (2)

where $z=(r_e-r_0)/a$ and r_0 denotes the distance at which n attains its maximum value n_0. The pitch-angle distribution P is assumed to be the same throughout the belt, and to have the form

$$P = A(\alpha) \sin^{\alpha+1}\theta,$$

where θ denotes the pitch-angle, α is a constant, and $A(\alpha)$ is a normalization factor. We shall

*Also engaged at the High Altitude Observatory, University of Colorado, in a program of research sponsored by the National Bureau of Standards and the Air Force Geophysical Directorate.

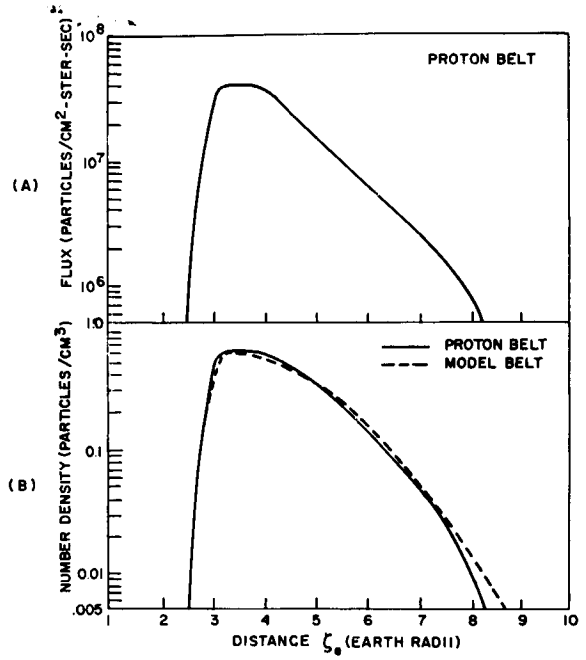


FIGURE 1.—(a) A typical flux distribution in the proton belt, as a function of radial distance from the earth's center; obtained in quiet conditions by Explorer XII; (b) The number density distribution in the model belt and in the observed proton belt.

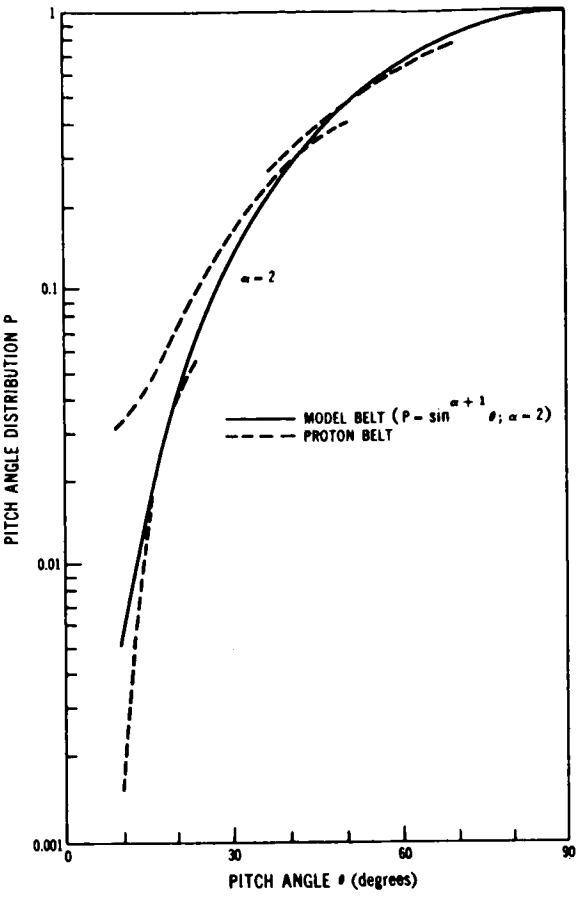


FIGURE 2.—The pitch-angle distribution in the model belt, together with the Explorer XII curves.

adopt the following numerical values:

$r_0 = 3.2a$; $\alpha = 2.0$;

$g_1 = 2.990$; $n_0 = 0.6/\text{cc}$;

$g_2 = 0.419$; $E = 500$ Kev.

The number density distribution with the

TABLE 1.—*Values for ΔF for Various Distances r_e from the Earth's Center (Equatorial Plane), where $g_1 = 2.990$, $g_2 = 0.419$, $\alpha = 2.0$ and $r_0 = 3.2a$*

Distance r_e (earth radii)	ΔF (gamma-cm³/Kev)	r_e (earth radii)	ΔF (gamma-cm³/Kev)	r_e (earth radii)	ΔF (gamma-cm³/Kev)
1.0	−0.128	3.0	−0.189	5.0	−0.203
1.2	−0.128	3.2	−0.227	5.5	−0.150
1.4	−0.127	3.4	−0.229	6.0	−0.086
1.6	−0.126	3.6	−0.233	6.5	−0.0243
1.8	−0.125	3.8	−0.237	7.0	0.025
2.0	−0.124	4.0	−0.240	7.5	0.056
2.2	−0.122	4.2	−0.240	8.0	0.071
2.4	−0.120	4.4	−0.237	8.5	0.059
2.6	−0.121	4.6	−0.230	9.0	0.045
2.8	−0.138	4.8	−0.218	10.0	0.023

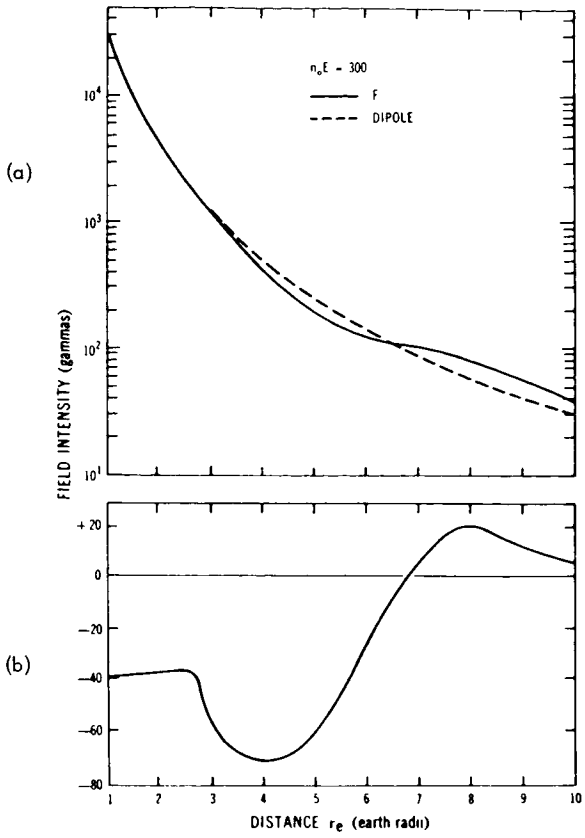


FIGURE 3.—(a) The distortion of the earth's dipole field produced by the model belt in the equatorial plane (b) and the distribution of the magnetic field ΔF produced by the model proton belt in the equatorial plane.

above values of g_1 and g_2 and corresponding to equations 1 and 2 is shown in figure 1 with the flux graph. In figure 2 the measured pitch-angle distributions (Davis, 1962) are compared with our distribution curve P. A more elaborate model belt could be used to fit other distributions of proton number density and velocity along an equatorial radius, but the numerical calculations would be more complicated. It is found, however, that for changes of the parameters g_1, g_2 and α, with the ranges 2.6–2.9, 0.38–0.42, and 2.0–3.0, respectively, the magnetic field produced by the model belt is not substantially affected.

THE MAGNETIC FIELD OF THE MODEL PROTON BELT

The magnetic field ΔF produced by our model belt is calculated for the above parameters by the method given in reference 2 and used in reference 3. As we ignore the field distortion caused by the belt, these results represent a first approximation, in which ΔF is proportional to the product n_0E. Let

$$\Delta F = n_0E\ \Delta F$$

where ΔF is given in gamma if n_0E is in units of Kev/cm^3. Note that in our model belt the value of n_0E is taken to be 300 Kev/cm^3. The values of ΔF are given in table 1 for a number of points along an equatorial radius.

Figure 3b shows, on a linear scale, the ΔF field along an equatorial radius. The intensity has a minimum value of about -72 gammas at $4.1a$, and a maximum value of about $+21$ gammas at $8.0a$. The corresponding field at the earth's surface is about -38 gammas.

Figure 3a gives, on a logarithmic scale, the distortion of the combined field of the earth's dipole and the model proton belt. Within $6.7a$ the model belt reduced the intensity of the earth's field and beyond this distance increased it. Such a distortion for quiet conditions can be checked by satellite-borne magnetometers, provided that other sources of distortion and higher order moments in the earth's field are known.*

ACKNOWLEDGMENTS

We wish to thank Leo R. Davis for discussing and communicating some of his results to us (figs. 1a and 2) prior to publication. The research reported in this paper was supported by a grant from the National Aeronautics and Space Administration.

REFERENCES

1. DAVIS, L. R., "Low-Energy Trapped Protons," in *GSFC Contributions to the COSPAR Meeting, May* 1962, NASA Technical Note D-1669.
2. AKASOFU, S.-I., and CHAPMAN, S., "Ring Current, Geomagnetic Disturbance and the Van Allen Radiation Belts," *J. Geophys. Res.*, **66**(5): 1321–1350, May 1961.
3. AKASOFU, S.-I., CAIN, J. C., and CHAPMAN, S., "The Magnetic Field of a Model Radiation Belt, Numerically Computed," NASA Technical Note D-1447, 1962; also, *J. Geophys. Res.*, **66**(12): 4013–4026, December 1961.
4. AKASOFU, S.-I., and CAIN, J. C., "The Magnetic Field of the Radiation Belts," *J. Geophys. Res.*, **67**(10): 4078–4080, September 1962.

*Further computations of the magnetic field of the radiation belts are available in reference 4.

THE MAGNETIC FIELD OF THE RADIATION BELTS*

SYUN-ICHI AKASOFU

University of Alaska

JOSEPH C. CAIN

Goddard Space Flight Center

The magnetic fields produced by various types of belts of trapped particles are presented in both tabular and graphical forms.

In this note we shall present in graphical form the magnetic fields produced by various types of belts of trapped particles. The chosen values of the parameters determining the field cover a wide range. It is hoped that, by interpolating the diagrams shown, the distortion of the earth's field can be approximately estimated for actually observed belts without a large amount of computation. References 1–5 discuss the method of computation.

Herein we shall state briefly the formulas used. The number density distribution along an equatorial radius r: $N=f_1(r)$ is given by:

$$N=N_0\exp(-g_1^2z^2)$$

($z<0$ for the inner part of the belt);

$$N=N_0\exp(-g_2^2z^2)$$

($z>0$ for the outer part of the belt),

where $z=(r-r_0)/a$, a is the radius of the earth, and r_0 denotes the distance at which N attains its maximum value N_0. Thus we may write

$$N=f_1(r_0, z, g_1, g_2).$$

The pitch-angle distribution, $P=f_2\ (\theta, \alpha)$, is

$$P=A(\alpha)\sin^{\alpha+1}\theta,$$

where θ denotes the pitch angle, α is a constant, and $A(\alpha)$ is a normalization factor.

In the energy spectrum $N=f_3(E)$, we shall consider only particles with a particular energy E (or speed v). The functions f_1, f_2, and f_3 determine the electric current intensity at any point in a dipole field. The ring current field ΔF is then calculated by a method shown in reference 2.

The ring current field ΔF is given by:

$$\Delta F=f(r, r_0, g_1, g_2, \alpha, N_0, E). \tag{1}$$

In the first approximation, ΔF is proportional to both N_0 and E. Therefore, equation 1 may be rewritten

$$\Delta F=\Delta f\ N_0\ E, \tag{2}$$

where

$$\Delta f=f'(r, r_0, g_1, g_2, \alpha). \tag{3}$$

We will show Δf, as a function of equatorial radial distance r, for various sets of r_0, g_1, g_2, and α. The vector ΔF is perpendicular to the equatorial plane and its value of ΔF is given in gammas, if the units of N_0 and E are chosen as cm^{-3} and Kev, respectively.

In figure 1(a) Δf is shown for different values of r_0 (app. A). The other parameters g_1, g_2, and α are chosen in such a way that they are suitable for the quiet-time proton belt (3):

$$g_1=2.990\ (N/N_0=1/10 \text{ at } z=-0.51),$$

$$g_2=0.419\ (N/N_0=1/10 \text{ at } z=+3.62),$$

$$\alpha=2.0.$$

The graph of N/N_0 is given in figure 1(b).

*This report has been published in substantially the same form minus appendix A, in *J. Geophys. Res.*, **67**(10): 4078–4080, September 1962.

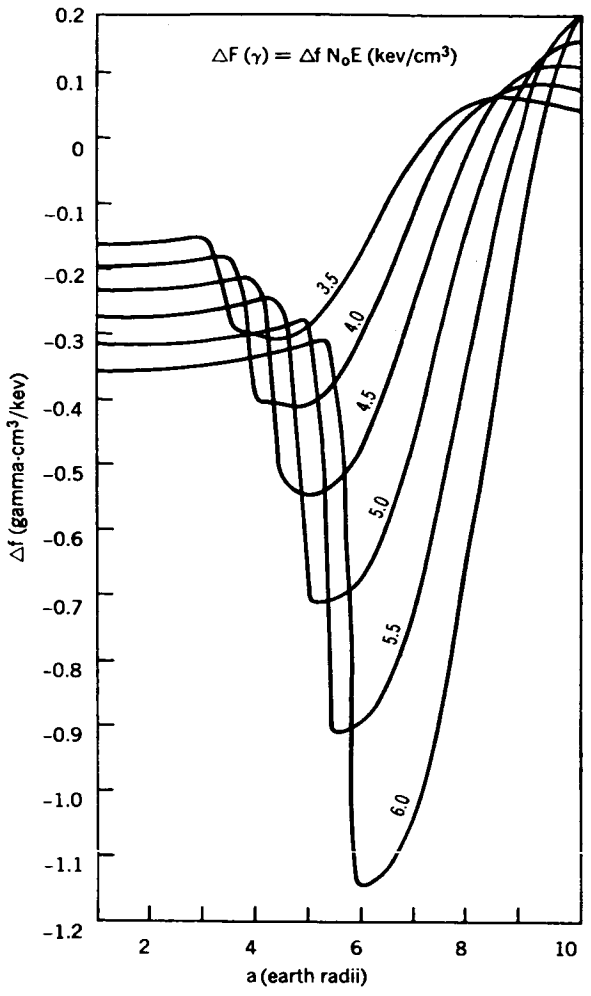


FIGURE 1(a).—Ring current field Δf for several values of r_0.

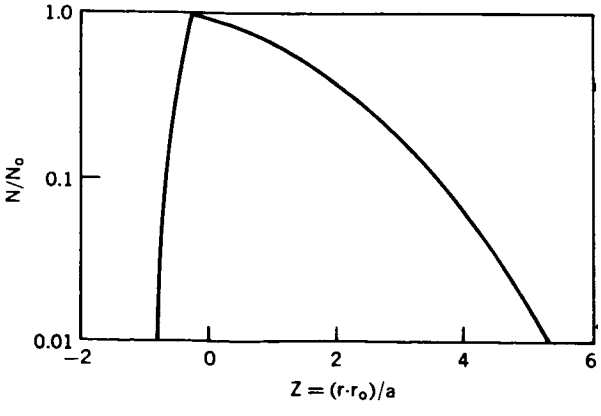


FIGURE 1(b).—Normalized electron density distribution.

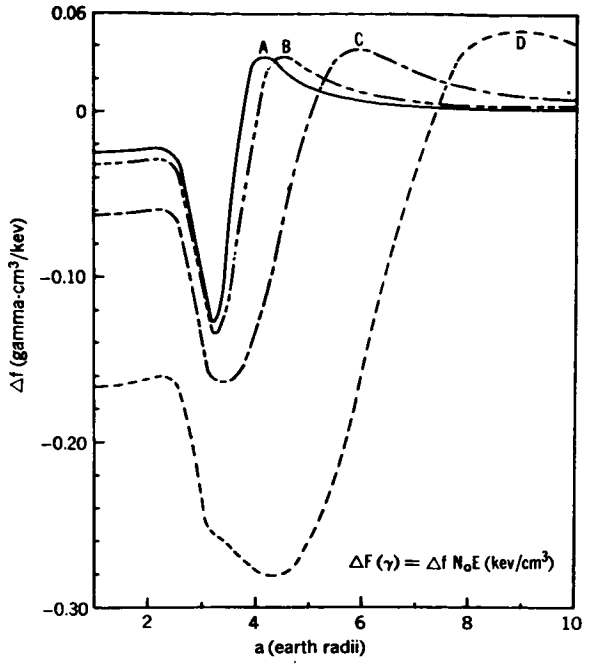


FIGURE 2(a).—Ring current field Δf for several values of outer slope. The values of g_2 for curves A, B, C, and D are 2.146, 1.517, 0.759, and 0.379, respectively.

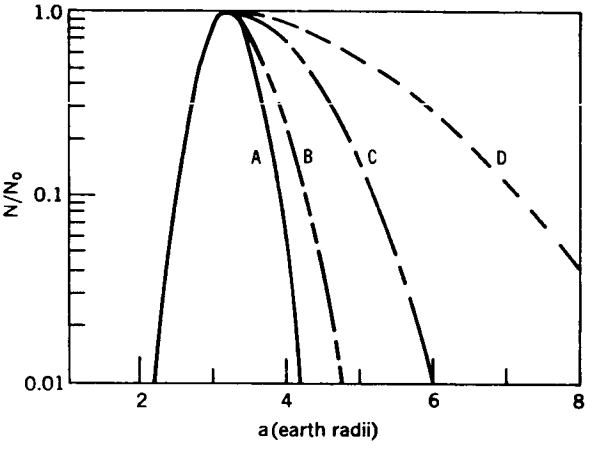


FIGURE 2(b).—The normalized electron density distribution. The curves A, B, C, and D are defined in figure 4a.

Figure 2(a) presents Δf for different values of g_2, namely:

$g_2 = 2.146$ ($N/N_0 = 1/100$ at $z = +1$),

$g_2 = 1.517$ ($N/N_0 = 1/10$ at $z = +1$),

$g_2 = 0.759$ ($N/N_0 = 1/10$ at $z = +2$),

$g_2 = 0.379$ ($N/N_0 = 1/10$ at $z = +4$).

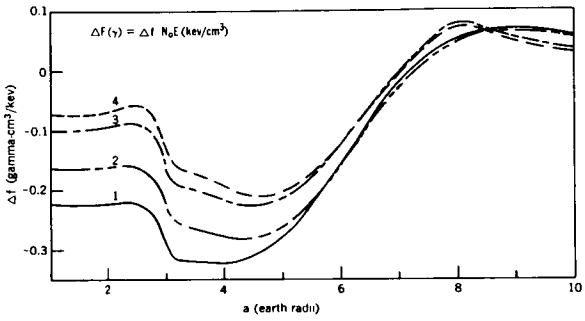


FIGURE 3(a).—Ring current field Δf for several α.

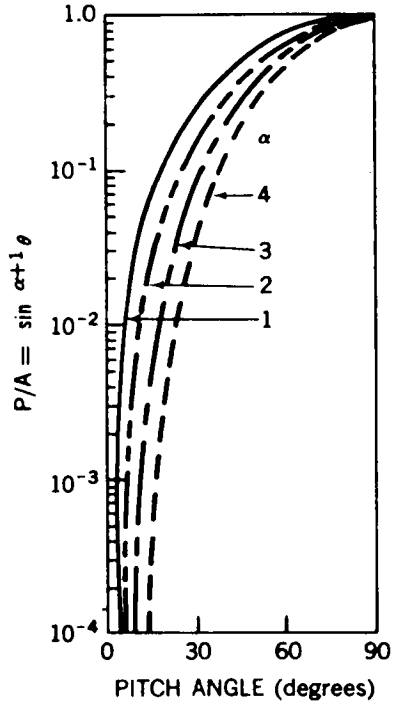


FIGURE 3(b).—Normalized pitch angle distribution for various α.

The graph of N/N_0 for these four cases is shown in figure 2(b). The other parameters are chosen as follows:

$$r_0 = 3.2a,$$

$$g_1 = 2.146 \ (N/N_0 = 1/100 \text{ at } z = -1),$$

$$\alpha = 2.0$$

In figure 3(a) Δf is given for different values of α; the graph of P/A for various values of α is given in figure 3(b). The other parameters are chosen as follows:

$$r_0 = 3.2a,$$

$$g_1 = 2.628 \ (N/N_0 = 1/1000 \text{ at } z = -1),$$

$$g_2 = 0.379 \ (N/N_0 = 1/10 \text{ at } z = +4).$$

The tabulated data from which figures 1(a), 2(a), and 3(a) are plotted are presented in appendix A.

REFERENCES

1. AKASOFU, S.-I., and CHAPMAN, S., "The Ring Current, Geomagnetic Disturbance and the Van Allen Radiation Belts," *J. Geophys. Res.*, **66**(5): 1321–1350, May 1961.
2. ———, CAIN, J. C., and CHAPMAN, S., "The Magnetic Field of a Radiation Belt, Numerically Computed," NASA Technical Note D–1447, November 1962; also *J. Geophys. Res.*, **66**(12): 4013–4020, Dec. 1961.
3. ———, ———, and ———, "The Magnetic Field of the Quiet-Time Proton Belt," NASA Technical Note D–1674, 1963; also *J. Geophys. Res.*, **67**(7): 2645–2647, July 1962.
4. BEARD, D. B., "Self-Consistent Calculation of the Ring Current," *J. Geophys. Res.*, **67**(9): 3615–3616, Aug. 1962.
5. AKASOFU, S.-I., "On a Self-Consistent Calculation of the Ring Current Field," *J. Geophys. Res.*, **67**(9): 3617–3618, Aug. 1962.

Appendix A

THE RING CURRENT FIELD Δf AS A FUNCTION OF r, g_2, AND α

TABLE A1.—*The Value of Δf as a Function of r_0*

r (earth radii)	Ring current field (gamma)									
	r_0 = 1.5a	r_0 = 2.0a	r_0 = 2.5a	r_0 = 3.0a	r_0 = 3.5a	r_0 = 4.0a	r_0 = 4.5a	r_0 = 5.0a	r_0 = 5.5a	r_0 = 6.0a
1.0	-0.0578	-0.0779	-0.1020	-0.1296	-0.1604	-0.1952	-0.2334	-0.2744	-0.3169	-0.3576
1.1	-0.0577	-0.0778	-0.1019	-0.1295	-0.1602	-0.1950	-0.2332	-0.2743	-0.3167	-0.3574
1.2	-0.0576	-0.0777	-0.1017	-0.1294	-0.1600	-0.1948	-0.2330	-0.2741	-0.3166	-0.3572
1.3	-0.0574	-0.0775	-0.1016	-0.1292	-0.1597	-0.1946	-0.2328	-0.2738	-0.3163	-0.3570
1.4	-0.0572	-0.0773	-0.1014	-0.1291	-0.1594	-0.1943	-0.2328	-0.2736	-0.3161	-0.3568
1.5	-0.0576	-0.0771	-0.1012	-0.1289	-0.1591	-0.1940	-0.2323	-0.2733	-0.3158	-0.3565
1.6	-0.0591	-0.0769	-0.1010	-0.1287	-0.1587	-0.1937	-0.2319	-0.2730	-0.3156	-0.3562
1.7	-0.0614	-0.0766	-0.1007	-0.1284	-0.1583	-0.1933	-0.2316	-0.2727	-0.3152	-0.3559
1.8	-0.0640	-0.0766	-0.1004	-0.1281	-0.1578	-0.1928	-0.2312	-0.2723	-0.3149	-0.3556
1.9	-0.0654	-0.0773	-0.1000	-0.1278	-0.1572	-0.1924	-0.2308	-0.2719	-0.3145	-0.3552
2.0	-0.0662	-0.0794	-0.0996	-0.1274	-0.1566	-0.1918	-0.2303	-0.2715	-0.3141	-0.3548
2.1	-0.0671	-0.0836	-0.0991	-0.1270	-0.1559	-0.1912	-0.2297	-0.2710	-0.3136	-0.3543
2.2	-0.0683	-0.0898	-0.0988	-0.1265	-0.1552	-0.1906	-0.2292	-0.2705	-0.3131	-0.3539
2.3	-0.0696	-0.0961	-0.0991	-0.1259	-0.1543	-0.1898	-0.2285	-0.2699	-0.3126	-0.3534
2.4	-0.0709	-0.0992	-0.1010	-0.1253	-0.1534	-0.1890	-0.2278	-0.2693	-0.3120	-0.3528
2.5	-0.0723	-0.1002	-0.1061	-0.1246	-0.1523	-0.1881	-0.2270	-0.2686	-0.3114	-0.3522
2.6	-0.0737	-0.1017	-0.1156	-0.1240	-0.1512	-0.1872	-0.2262	-0.2678	-0.3107	-0.3515
2.7	-0.0751	-0.1033	-0.1288	-0.1237	-0.1502	-0.1861	-0.2253	-0.2870	-0.3099	-0.3508
2.8	-0.0764	-0.1051	-0.1415	-0.1247	-0.1498	-0.1849	-0.2243	-0.2661	-0.3091	-0.3501
2.9	-0.0776	-0.1070	-0.1472	-0.1288	-0.1517	-0.1836	-0.2231	-0.2651	-0.3083	-0.3493
3.0	-0.0786	-0.1089	-0.1484	-0.1389	-0.1590	-0.1822	-0.2219	-0.2641	-0.3073	-0.3484
3.1	-0.0794	-0.1107	-0.1502	-0.1569	-0.1761	-0.1807	-0.2206	-0.2629	-0.3063	-0.3474
3.2	-0.0800	-0.1123	-0.1522	-0.1810	-0.2061	-0.1795	-0.2192	-0.2617	-0.3052	-0.3464
3.3	-0.0804	-0.1138	-0.1544	-0.2034	-0.2452	-0.1794	-0.2177	-0.2604	-0.3040	-0.3454
3.4	-0.0804	-0.1150	-0.1567	-0.2130	-0.2807	-0.1831	-0.2160	-0.2589	-0.3028	-0.3442
3.5	-0.0801	-0.1160	-0.1589	-0.2142	-0.2949	-0.1953	-0.2142	-0.2874	-0.3014	-0.3430
3.6	-0.0794	-0.1165	-0.1609	-0.2160	-0.2955	-0.2230	-0.2124	-0.2557	-0.2999	-0.3416
3.7	-0.0784	-0.1167	-0.1626	-0.2183	-0.2969	-0.2700	-0.2110	-0.2539	-0.2983	-0.3402
3.8	-0.0771	-0.1164	-0.1639	-0.2207	-0.2989	-0.3304	-0.2114	-0.2520	-0.2967	-0.3387
3.9	-0.0753	-0.1156	-0.1648	-0.2231	-0.3011	-0.3842	-0.2176	-0.2499	-0.2948	-0.3371
4.0	-0.0732	-0.1143	-0.1652	-0.2252	-0.3031	-0.4054	-0.2367	-0.2476	-0.2929	-0.3353
4.1	-0.0707	-0.1126	-0.1651	-0.2270	-0.3049	-0.4056	-0.2784	-0.2455	-0.2908	-0.3335
4.2	-0.0679	-0.1103	-0.1643	-0.2284	-0.3062	-0.4064	-0.3480	-0.2440	-0.2886	-0.3315
4.3	-0.0648	-0.1075	-0.1628	-0.2292	-0.3067	-0.4078	-0.4361	-0.2453	-0.2862	-0.3294
4.4	-0.0614	-0.1041	-0.1607	-0.2293	-0.3065	-0.4094	-0.5135	-0.2550	-0.2837	-0.3272
4.5	-0.0578	-0.1004	-0.1579	-0.2286	-0.3053	-0.4107	-0.5439	-0.2832	-0.2810	-0.3248
4.6	-0.0540	-0.0961	-0.1543	-0.2271	-0.3031	-0.4115	-0.5437	-0.3431	-0.2785	-0.3223
4.7	-0.0500	-0.0915	-0.1501	-0.2247	-0.2997	-0.4115	-0.5435	-0.4414	-0.2770	-0.3196
4.8	-0.0458	-0.0864	-0.1452	-0.2215	-0.2952	-0.4105	-0.5439	-0.5643	-0.2796	-0.3168
4.9	-0.0416	-0.0811	-0.1397	-0.2173	-0.2896	-0.4084	-0.5443	-0.6710	-0.2939	-0.3137
5.0	-0.0373	-0.0754	-0.1335	-0.2121	-0.2827	-0.4050	-0.5442	-0.7131	-0.3337	-0.3106
5.5	-0.0167	-0.0453	-0.0958	-0.1733	-0.2315	-0.3661	-0.5265	-0.7064	-0.9150	-0.3853
6.0	-0.0002	-0.0168	-0.0532	-0.1190	-0.1601	-0.2916	-0.4658	-0.6720	-0.8990	-1.1490
6.5	+0.0103	+0.0050	-0.0150	-0.0610	-0.0838	-0.1944	-0.3619	-0.5831	-0.8426	-1.1220
7.0	+0.0153	+0.0181	+0.0130	-0.0109	-0.0175	-0.0946	-0.2330	-0.4427	-0.7182	-1.0370
7.5	+0.0162	+0.0237	+0.0290	+0.0243	+0.0296	-0.0106	-0.1050	-0.2757	-0.5335	-0.8692
8.0	+0.0149	+0.0242	+0.0352	+0.0436	+0.0560	+0.0475	-0.0002	-0.1144	-0.3211	-0.6314
8.5	+0.0128	+0.0219	+0.0348	+0.0502	+0.0654	+0.0787	+0.0705	+0.0148	-0.1209	-0.3656
9.0	+0.0106	+0.0187	+0.0312	+0.0487	+0.0641	+0.0889	+0.1073	+0.1000	+0.0365	-0.1200
9.5	+0.0087	+0.0155	+0.0266	+0.0434	+0.0573	+0.0859	+0.1182	+0.1430	+0.1384	+0.0700
10.0	+0.0071	+0.0128	+0.0222	+0.0370	+0.0490	+0.0766	+0.1131	+0.1545	+0.1885	+0.1910

TABLE A2.—*The Values of Δf as a Function of g_2*

r (earth radii)	Ring current field (gamma)				
	$g_2 = 2.146$	$g_2 = 1.517$	$g_2 = 0.759$	$g_2 = 0.509$	$g_2 = 0.379$
1.0	-0.0245	-0.0320	-0.0638	-0.1075	-0.1663
1.1	-0.0245	-0.0319	-0.0637	-0.1073	-0.1661
1.2	-0.0243	-0.0318	-0.0635	-0.1071	-0.1658
1.3	-0.0242	-0.0317	-0.0633	-0.1068	-0.1655
1.4	-0.0241	-0.0315	-0.0631	-0.1065	-0.1651
1.5	-0.0239	-0.0313	-0.0628	-0.1062	-0.1647
1.6	-0.0237	-0.0311	-0.0625	-0.1058	-0.1642
1.7	-0.0235	-0.0309	-0.0621	-0.1053	-0.1636
1.8	-0.0233	-0.0306	-0.0617	-0.1048	-0.1630
1.9	-0.0230	-0.0302	-0.0613	-0.1042	-0.1623
2.0	-0.0227	-0.0299	-0.0607	-0.1036	-0.1616
2.1	-0.0224	-0.0296	-0.0603	-0.1029	-0.1608
2.2	-0.0223	-0.0294	-0.0599	-0.1024	-0.1601
2.3	-0.0226	-0.0297	-0.0600	-0.1023	-0.1599
2.4	-0.0240	-0.0310	-0.0610	-0.1032	-0.1606
2.5	-0.0273	-0.0342	-0.0640	-0.1059	-0.1631
2.6	-0.0339	-0.0407	-0.0702	-0.1119	-0.1688
2.7	-0.0449	-0.0516	-0.0808	-0.1222	-0.1789
2.8	-0.0610	-0.0676	-0.0964	-0.1375	-0.1940
2.9	-0.0812	-0.0876	-0.1160	-0.1568	-0.2131
3.0	-0.1024	-0.1087	-0.1366	-0.1771	-0.2331
3.1	-0.1199	-0.1259	-0.1534	-0.1935	-0.2491
3.2	-0.1282	-0.1341	-0.1611	-0.2008	-0.2560
3.3	-0.1237	-0.1335	-0.1630	-0.2028	-0.2579
3.4	-0.1061	-0.1261	-0.1640	-0.2051	-0.2604
3.5	-0.0786	-0.1123	-0.1637	-0.2073	-0.2634
3.6	-0.0469	-0.0931	-0.1618	-0.2093	-0.2665
3.7	-0.0171	-0.0706	-0.1580	-0.2107	-0.2697
3.8	+0.0066	-0.0470	-0.1523	-0.2114	-0.2727
3.9	+0.0225	-0.0245	-0.1447	-0.2112	-0.2755
4.0	+0.0310	-0.0050	-0.1352	-0.2100	-0.2779
4.1	+0.0339	+0.0106	-0.1240	-0.2077	-0.2797
4.2	+0.0334	+0.0218	-0.1114	-0.2042	-0.2810
4.3	+0.0312	+0.0290	-0.0978	-0.1995	-0.2815
4.4	+0.0283	+0.0328	-0.0835	-0.1935	-0.2813
4.5	+0.0255	+0.0341	-0.0688	-0.1864	-0.2802
4.6	+0.0228	+0.0335	-0.0543	-0.1781	-0.2783
4.7	+0.0205	+0.0319	-0.0402	-0.1688	-0.2754
4.8	+0.0185	+0.0298	-0.0269	-0.1585	-0.2716
4.9	+0.0168	+0.0274	-0.0145	-0.1474	-0.2668
5.0	+0.0152	+0.0251	-0.0034	-0.1356	-0.2610
5.5	+0.0100	+0.0162	+0.0315	-0.0719	-0.2186
6.0	+0.0070	+0.0111	+0.0379	-0.0143	-0.1592
6.5	+0.0051	+0.0080	+0.0319	+0.0247	-0.0940
7.0	+0.0039	+0.0060	+0.0242	+0.0432	-0.0343
7.5	+0.0030	+0.0047	+0.0182	+0.0466	+0.0118
8.0	+0.0024	+0.0037	+0.0140	+0.0421	+0.0415
8.5	+0.0020	+0.0030	+0.0110	+0.0351	+0.0561
9.0	+0.0016	+0.0025	+0.0089	+0.0285	+0.0596
9.5	+0.0013	+0.0020	+0.0073	+0.0230	+0.0565
10.0	+0.0011	+0.0017	+0.0061	+0.0188	+0.0502

TABLE A3.—*The Values of Δf as a Function of α*

r (earth radii)	Ring current field (gamma)			
	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 3.0$	$\alpha = 4.0$
1.0	-0.2240	-0.1649	-0.1017	-0.0784
1.1	-0.2239	-0.1647	-0.1013	-0.0779
1.2	-0.2239	-0.1644	-0.1008	-0.0773
1.3	-0.2239	-0.1641	-0.1002	-0.0766
1.4	-0.2238	-0.1638	-0.0996	-0.0758
1.5	-0.2237	-0.1634	-0.0989	-0.0749
1.6	-0.2237	-0.1629	-0.0981	-0.0739
1.7	-0.2235	-0.1624	-0.0972	-0.0728
1.8	-0.2234	-0.1618	-0.0962	-0.0715
1.9	-0.2232	-0.1611	-0.0950	-0.0700
2.0	-0.2230	-0.1604	-0.0938	-0.0684
2.1	-0.0228	-0.1596	-0.0923	-0.0665
2.2	-0.0226	-0.1587	-0.0908	-0.0645
2.3	-0.2224	-0.1578	-0.0891	-0.0623
2.4	-0.2226	-0.1572	-0.0877	-0.0603
2.5	-0.2238	-0.1576	-0.0873	-0.0592
2.6	-0.2273	-0.1605	-0.0894	-0.0606
2.7	-0.2350	-0.1679	-0.0962	-0.0669
2.8	-0.2489	-0.1821	-0.1102	-0.0805
2.9	-0.2692	-0.2036	-0.1319	-0.1023
3.0	-0.2929	-0.2292	-0.1582	-0.1290
3.1	-0.3126	-0.2511	-0.1811	-0.1525
3.2	-0.3197	-0.2599	-0.1908	-0.1626
3.3	-0.3193	-0.2611	-0.1926	-0.1648
3.4	-0.3196	-0.2631	-0.1954	-0.1681
3.5	-0.3202	-0.2656	-0.1988	-0.1721
3.6	-0.3210	-0.2685	-0.2027	-0.1767
3.7	-0.3217	-0.2714	-0.2067	-0.1815
3.8	-0.3223	-0.2742	-0.2107	-0.1863
3.9	-0.3225	-0.2768	-0.2146	-0.1911
4.0	-0.3222	-0.2790	-0.2182	-0.1956
4.1	-0.3214	-0.2807	-0.2213	-0.1998
4.2	-0.3200	-0.2819	-0.2239	-0.2034
4.3	-0.3178	-0.2823	-0.2258	-0.2065
4.4	-0.3149	-0.2820	-0.2270	-0.2088
4.5	-0.3111	-0.2809	-0.2274	-0.2104
4.6	-0.3064	-0.2789	-0.2269	-0.2111
4.7	-0.3009	-0.2759	-0.2255	-0.2109
4.8	-0.2944	-0.2721	-0.2232	-0.2098
4.9	-0.2871	-0.2672	-0.2198	-0.2076
5.0	-0.2788	-0.2614	-0.2155	-0.2045
5.5	-0.2254	-0.2189	-0.1794	-0.1740
6.0	-0.1581	-0.1594	-0.1244	-0.1236
6.5	-0.0881	-0.0941	-0.0614	-0.0637
7.0	-0.0263	-0.0344	-0.0020	-0.0060
7.5	+0.0202	+0.0117	+0.0455	+0.0409
8.0	+0.0493	+0.0414	+0.0774	+0.0731
8.5	+0.0630	+0.0560	+0.0672	+0.0634
9.0	+0.0655	+0.0596	+0.0506	+0.0473
9.5	+0.0615	+0.0564	+0.0394	+0.0366
10.0	+0.0545	+0.0501	+0.0315	+0.0291

REVIEW OF DIRECT MEASUREMENTS OF INTERPLANETARY DUST FROM SATELLITES AND PROBES

W. M. ALEXANDER, C. W. McCRACKEN, L. SECRETAN, AND O. E. BERG

Goddard Space Flight Center

Interplanetary dust particle experiments have been flown on various types of spacecraft. Measurements of dust-particle momentum, kinetic energy, and impact damage have been obtained from a variety of sensors.

More than 10,000 dust-particle impacts have been measured by acoustical and light-flash detectors. A cumulative mass distribution curve valid for average conditions in the vicinity of the earth has been derived from the direct measurements for dust particles with masses between about 10^{-13} and 10^{-6} gm.

The results of direct measurements from various penetration and fracture experiments are presented here. Comparison of these data with the average mass distribution curve shows no major discrepancies among the measurements from all of the different sensor systems used. Two satellite experiments, Vanguard III (1959?) and Explorer I (1958α), give evidence for dust-particle streams. Vanguard III measured 2800 events in a 70-hour interval coincident in time with the expected return of the Leonid meteor shower. The average influx rate during the dust-particle shower was more than one order of magnitude greater than the nonshower influx rate measured with the same satellite. There were rapid fluctuations during the shower when the influx rates increased by as much as three orders of magnitude above the nonshower average rate. For a 10-hour period, Explorer I detected a dust particle shower with impact rates as high as 25 times the average impact rate during the satellite's lifetime.

The distribution curve obtained with the direct measurements differs from that expected on the basis of extrapolations of meteor observations. Small dust particles dominate the accretion by the earth of interplanetary matter; the accretion rate is of the order of 10^4 tons per day.

INTRODUCTION

An important constituent of the solar system is the cloud of dust surrounding the sun. Knowledge concerning the origin, composition, and dynamic properties of these dust particles is fundamental to considerations of the solar system.

Various techniques have been used in studying the distributions and physical properties of dust particles in the zodiacal cloud. These include: ground-based visual, photographic, and radar metoer observations; photometric observations of the zodiacal light and solar F corona; and various types of collection techniques and laboratory analyses.

From observations of collisions of individual meteoroids with the earth's atmosphere, vector velocity, mass, density, and composition have been measured. These studies are limited, in most cases, to particle masses greater than 10^{-4} gm. Additional information has been obtained from the collection of remnants of meteoroidal and dust-particle material which have survived passage through the atmosphere. Composition, age, and gross-accretion rates have also been determined from these analyses. Zodiacal light observations represent the only ground-based technique which has been used to measure the size distribution of interplanetary dust particles with masses less than 10^{-6} gm.

Another technique for measuring the physical characteristics and dynamic properties of dust particles became a reality with the advent of vehicles capable of placing experiments in space.

Detectors have been developed which are capable of measurements on individual dust particles with masses as small as 10^{-13} gm. The dust-particle measurements in space started with experiments borne by sounding rockets in 1949. Subsequent to the successful launching of Explorer I (1958α), results have been obtained from experiments on 14 United States and Soviet Union satellites and space probes. Six different types of sensors have been used to measure the following particle parameters: Momentum, kinetic energy, penetration, and fracture properties. The results of these rocket and satellite measurements are presented in the following sections.

DIRECT MEASUREMENTS FROM MICROPHONE SYSTEMS

The microphone type of dust particle sensor has basically consisted of a piezoelectric crystal microphone attached to a metallic sounding board. The electrical signal generated as an impacting dust particle delivers a mechanical impulse to the sounding board is amplified and pulse-height analyzed in order to obtain information about the particle. Analog calibrations, performed in the laboratory by dropping carefully selected glass spheres onto the sounding boards, have consistently shown (for low velocities) that the microphone system is sensitive to the momentum of an impacting particle. Hypervelocity studies with microparticles from shaped charges[1] tend to confirm that the microphone systems are also sensitive to the momentum of a particle impacting at hypervelocity. An energy dependence can be obtained by the use of the theoretical results of Stanyukovich.[2] Lavrentyev[3] theoretically finds that the sensitivity lies somewhere between momentum and energy.

The direct measurements can presently be expressed in terms of the particle mass, subject only to minor uncertainties. These uncertainties include the choice of an average speed (relative to the satellite) for the particles; the determination of an effective coefficient of restitution for hypervelocity microparticle impacts; and the computation of the appropriate correction factors for shielding by the earth, for the solid viewing angle of the sensor, and for the orientation of the solid viewing angle relative to the apex of the earth's motion. The various correction factors have been chosen in such a way that the minor corrections which will most probably become necessary will leave the results of this analysis essentially unchanged.

Microphone systems have provided the greatest quantity of information about the interplantary dust particles. These systems have flown on more vehicles and over a greater range of geocentric distances than any of the other systems. In addition, they are more nearly calibrated than other types of dust particle sensors. It will be assumed, for the present analysis, that the microphone systems are momentum-sensitive and that the ratio of the mechanical impulse to the impact momentum of a particle is unity. A small correction, estimated to be a factor of 2 or 3, can be introduced later when the appropriate hypervelocity laboratory studies have been completed.

The microphone system on Explorer VIII (1960ξ) used two metallic sounding boards attached to a conical section of the spin-stabilized satellite. The solid viewing angle of the system was 2π steradians and remained almost centered on the antapex of the earth's motion during the

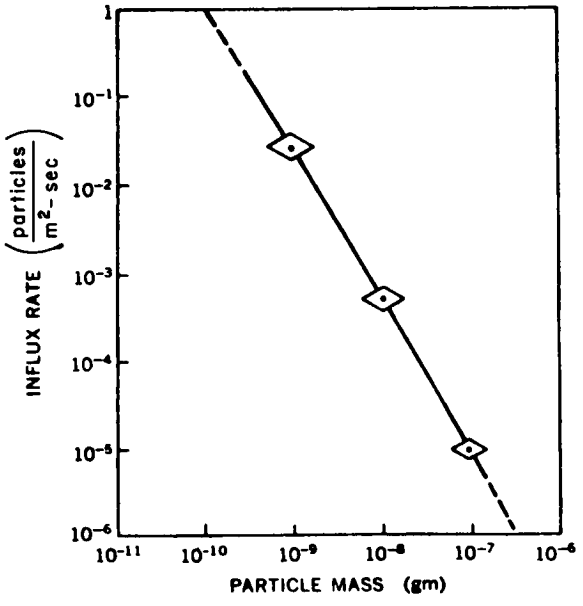


FIGURE 1.—Cumulative mass distribution established by the microphone system on Explorer VIII. The largest symbol represents the data point of highest significance, and the diminished size of the symbols indicates lower significance.

TABLE 1.—*Direct Measurements Obtained with Microphone Systems on U.S. Satellites and Space Probes*

Spacecraft	Momentum sensitivity (dyne-sec)	Mass sensitivity (gm)	Effective area (m^2)	Exposure time (sec)	Exposure (m^2-sec)	Number of particles	Cumulative influx rate (particles/m^2-sec)	
							Observed	Corrected
Explorer VIII	2.5×10^{-3}–2.5×10^{-2}	1.0×10^{-9}–1.0×10^{-8}	7.0×10^{-2}	3.5×10^{6}	2.4×10^{5}	~3650	1.5×10^{-2}	3.0×10^{-2}
	2.5×10^{-2}–2.5×10^{-1}	1.0×10^{-8}–1.0×10^{-7}				~75	3.1×10^{-4}	6.2×10^{-4}
	$>2.5\times10^{-1}$	$>1.0\times10^{-7}$				1 or 2	~5.0×10^{-6}	~1.0×10^{-5}
Vanguard III	$>1.0\times10^{-2}$	$>3.3\times10^{-9}$	4.0×10^{-1}	6.9×10^{6}	2.8×10^{6}	~3500	1.3×10^{-3}	2.0×10^{-3}
Explorer I	$>2.5\times10^{-3}$	$>8.3\times10^{-10}$	2.3×10^{-1}	7.9×10^{4}	1.8×10^{4}	145	8.4×10^{-3}	1.7×10^{-2}
Pioneer I	$>1.5\times10^{-4}$	$>5.0\times10^{-11}$	3.9×10^{-2}	1.1×10^{5}	4.2×10^{3}	17	4.0×10^{-3}	1.6×10^{-2}
Ranger I	$>3.0\times10^{-5}$	$>1.0\times10^{-11}$	8.0×10^{-4}	1.1×10^{4}	8.8	64	7.3	4.0×10
Midas II	$>3.0\times10^{-4}$	$>1.0\times10^{-10}$	6.9×10^{-2}	4.0×10^{3}	2.7×10^{2}	67	2.5×10^{-1}	5.0×10^{-1}
Samos II	$\gtrsim3.0\times10^{-4}$	$\gtrsim1.0\times10^{-10}$	6.9×10^{-2}	(?)	(?)	(?)	3.4×10^{-1}	6.8×10^{-1}
SLV–1	$\gtrsim9.0\times10^{-3}$	$\gtrsim3.0\times10^{-9}$	8.0×10^{-1}	9.5×10^{2}	7.6×10^{2}	10	1.3×10^{-2}	2.6×10^{-2}

lifetime of the experiment. From an analog calibration performed in the laboratory, the limiting sensitivities for the three ranges of sensitivity were found to be 2.5×10^{-3}, 2.5×10^{-2}, and 2.5×10^{-1} dyne second. These may be expressed in terms of particle mass as 1.0×10^{-9}, 1.0×10^{-8}, and 1.0×10^{-7} gm, respectively, if an average speed (relative to the satellite) of 25 km/sec is assumed. The large separations of the limiting sensitivites used in the Explorer VIII microphone system are of particular importance in view of the excellent data sample which was obtained. The magnitude of the total range of sensitivity allows not only a definition of the influx rates within the three ranges but also the establishment of the shape of a segment of an average mass distribution curve.

The fine structure in the influx rate measured with the Explorer VIII microphone system is presently being analyzed. The large variations (at least plus or minus an order of magnitude from the mean, within intervals of a few hours, for particle masses of about 10^{-9} gm) are being studied to determine their physical significance. The preliminary readouts of the total numbers of impacts used in establishing the average mass distribution curves that were reported earlier[4,5] have been confirmed. Exact specification of these numbers awaits completion of the analyses which are in progress, but "revised preliminary" numbers may be given. They are: 3726 dust particles with momenta of 2.5×10^{-3} dyne-sec and greater, 76 with momenta of 2.5×10^{-2} dyne-sec and greater, and 1 or 2 with momenta of 2.5×10^{-1} dyne-sec and greater. (The corresponding limits on particle mass were given above.) The area of the sounding boards was 7.0×10^{-2} m^2, and the lifetime of the experiment was 3.47×10^{6} seconds.

A correction factor of 2 is applied to convert the influx rates to omnidirectional values before the data are plotted as the cumulative mass distribution curve shown in figure 1. Although the data point for the scale of lowest sensitivity is not very significant, it lies on the straight-line segment indicated by the two data points which are in figure 1 is:

$$\log I = -17.0 - 1.70 \log m, \qquad (1)$$

where I is the cumulative omnidirectional influx rate in particles/m^2-sec, and m is the particle mass in grams.

The spacecraft from which direct measurements with microphone systems have been obtained in the United States are listed, together with the relevant data, in table 1. Average particle speeds of 30 km/sec have been used for the omnidirectional sensors and for sensors mounted on spinning or tumbling vehicles which viewed most of the celestial sphere. An attempt has been made to apply correction factors for shielding by the earth and lack of omnidirectionality of the sensors. All data were converted to omnidirectional influx rates.

Reading of the telemetered data from Vanguard III (1959η) has now been completed and is in the stages of analysis. More than 6000 impacts were recorded during the 80-day lifetime of the experiment. Of this number, approximately 2800 impacts occurred in a 70-hour interval on November 16 to 18, which coincides in time with the expected annual return of the Leonid meteor shower. An average influx rate has been computed from the Vanguard III data on the basis of approximately 3500 impacts. A factor of 1.5 was used to correct for shielding by the earth.

The data given for Explorer I and Pioneer I are those reported by Dubin.[6,7] The total number of impacts (145) for Explorer I was used in computing an average influx rate even though more than half of the impacts probably represented an interplanetary dust particle event.[8,9] The high influx rates during this event were nearly counterbalanced by an interval of low rates, so an influx rate computed from the total number of impacts serves very well in the present analysis. A factor of 2 was used in correcting for shielding by the earth and in allowing for the fact that the sensor was not completely omnidirectional. The microphone system on Pioneer I registered 25 impacts, of which 17 are considered to represent impacting dust particles. No correction for shielding by the earth was made, because Pioneer I spent most of its time at large geocentric distances (2 to 19 earth radii). A factor of 4 was used in converting to an omnidirectional influx rate.

A preliminary readout of the data from the meirophone and coated photomultiplier systems on Ranger I (1961ϕ) has been reported by

TABLE 2.—*Direct Measurements Obtained with Microphone Systems on the OSU Rockets*

Rocket	Momentum sensitivity (dyne-sec)	Particle speed (km/sec)	Mass sensitivity (gm)	Number of impacts	Exposure $h>110$ km (m^2-sec)	Cumulative influx rate (particles/ m^2-sec)
Aerobee No. 80	$>6.0\times10^{-4}$	70	$>8.6\times10^{-11}$	49	5.0	9.8
	$>3.0\times10^{-3}$	70	$>4.3\times10^{-10}$	10	5.0	2.0
	$>1.0\times10^{-3}$	40	$>2.5\times10^{-10}$	3	50.0	6.0×10^{-2}
	$>3.0\times10^{-3}$	40	$>7.5\times10^{-10}$	1	50.0	2.0×10^{-2}
Aerobee No. 88	$>1.3\times10^{-4}$	20	$>6.5\times10^{-11}$	6	3.0	2.0
	$>2.0\times10^{-3}$	20	$>1.0\times10^{-9}$	1	3.0	3.3×10^{-1}
	$>4.7\times10^{-4}$	35	$>1.3\times10^{-10}$	17	30.0	5.7×10^{-1}
	$>1.0\times10^{-3}$	35	$>2.9\times10^{-10}$	7	30.0	2.3×10^{-1}
Nike–Cajun AF–2	$>6.0\times10^{-4}$	40	$>1.5\times10^{-10}$	45	31.0	1.5
	$>1.2\times10^{-3}$		$>3.0\times10^{-10}$	15		4.8×10^{-1}
	$>4.0\times10^{-3}$		$>1.0\times10^{-9}$	3		9.7×10^{-2}
Nike–Cajun AA6.203	$>3.0\times10^{-4}$	35	$>8.6\times10^{-11}$	55	37.0	1.5
	$>3.0\times10^{-3}$		$>8.6\times10^{-10}$	3		8.1×10^{-2}
Nike–Cajun AA6.204	$>7.0\times10^{-4}$	40	$>1.8\times10^{-10}$	32	33.0	9.7×10^{-1}
	$>3.0\times10^{-3}$		$>7.5\times10^{-10}$	1		3.0×10^{-2}
Nike–Cajun AA6.206	$>1.5\times10^{-4}$	35	$>4.3\times10^{-11}$	12	24.0	5.0×10^{-1}
	$>1.0\times10^{-3}$		$>2.9\times10^{-10}$	1		4.2×10^{-2}
	$>7.0\times10^{-4}$		$>2.0\times10^{-10}$	6		2.5×10^{-1}
Spaerobee 10.01	$>5.0\times10^{-4}$	60	$>8.3\times10^{-11}$	20	8.1	2.5

TABLE 3.—*Direct Measurements Obtained with Microphone Systems on the Soviet Union Rockets, Satellites, and Space Probes*

Vehicle	Mass sensitivity (gm)		Effective area (m^2)	Exposure time (sec)	Exposure (m^2-sec)	Number of particles	Influx rate particles/m^2-sec)	
	$v=40$ km/sec	$v=30$ km/sec					(Nazarova)	Cumulative
Sputnik III	8.0×10^{-9}–2.7×10^{-8}	1.4×10^{-8}–4.8×10^{-8}	0.34	$\sim8.0\times10^{5}$	3.0×10^{5}	(?)	(See text)	$<1.0\times10^{-4}$
	2.7×10^{-8}–1.5×10^{-7}	4.8×10^{-8}–2.7×10^{-7}						
	1.5×10^{-7}–5.6×10^{-6}	2.7×10^{-7}–1.0×10^{-5}						
	$>5.6\times10^{-6}$	$>1.0\times10^{-5}$						
Lunik I	2.5×10^{-9}–1.5×10^{-8}	4.4×10^{-9}–2.7×10^{-8}	0.2	3.6×10^{4}	7.2×10^{3}	<16	$<2.0\times10^{-3}$	$<2.9\times10^{-3}$
	1.5×10^{-8}–2.0×10^{-7}	2.7×10^{-8}–3.6×10^{-7}				<4	$<5.0\times10^{-4}$	$<7.0\times10^{-4}$
	$>2.0\times10^{-7}$	$>3.6\times10^{-7}$				<1	$<1.0\times10^{-4}$	$<1.4\times10^{-4}$
Lunik II	2.0×10^{-9}–6.0×10^{-9}	3.6×10^{-9}–1.1×10^{-8}	0.2	1.1×10^{5}	2.2×10^{4}	0	$<5.0\times10^{-5}$	
	6.0×10^{-9}–1.5×10^{-8}	1.1×10^{-8}–2.7×10^{-8}				0	$<5.0\times10^{-5}$	
	$>1.5\times10^{-8}$	$>2.7\times10^{-8}$				2	9.0×10^{-5}	9.1×10^{-5}
Interplanetary station	1.0×10^{-9}–3.0×10^{-9}	1.8×10^{-9}–5.3×10^{-9}	0.1	2.3×10^{4}	2.3×10^{3}	1	4.0×10^{-4}	3.0×10^{-3}
	3.0×10^{-9}–8.0×10^{-9}	5.3×10^{-9}–1.4×10^{-8}				5	2.0×10^{-3}	2.6×10^{-3}
	$>8.0\times10^{-9}$	$>1.4\times10^{-8}$				1	4.0×10^{-4}	4.3×10^{-4}
Geophysical Rocket I		$\gtrsim2.5\times10^{-9}$	4.0	1.3×10^{2}	5.4×10^{2}	(?)	6.0×10^{-2}	6.0×10^{-2}
Geophysical Rocket II		$\gtrsim2.5\times10^{-9}$	4.0	1.5×10^{2}	5.9×10^{2}	(?)	5.0×10^{-2}	5.0×10^{-2}
Geophysical Rocket III		$\gtrsim2.5\times10^{-9}$	4.0	8.5×10	3.4×10^{2}	(?)	7.5×10^{-1}	7.5×10^{-1}

Alexander and Berf.[10] In this system, the two sensors were capable of operating in coincidence as well as independently. The data for the microphone system alone are given in table 1.

The results from the microphone system on Midas II (1960ζ1) and preliminary results from a similar system on Samos II (1961α1) have been reported by Soberman and Della Lucca.[11] The data were obtained—as in the case of Explorer I—in real time as the satellite passed over telemetry stations.

The data from a microphone system on SLV-1 (a Vanguard satellite that failed to achieve orbit) have been reported by LaGow, Schaefer, and Schaffert.[12] A data point which is of use in the present analysis can be obtained if the sensitivity of the system is computed on the basis of momentum rather than energy. The bursts of counts observed were most likely produced when the expended third stage motor sputtered and bumped the separated payload (LaGow, private communication). Therefore, only 10 of the 17 impacts are used in computing an influx rate. The value of the mass sensitivity assumed for this system has been computed from calibration data which were given.[12]

Some of the earliest direct measurements of quantitative value were obtained with microphone systems on a series of seven successful high-altitude rockets instrumented and flown by Oklahoma State University, hereafter referred to as OSU.[5,13,14] The data from these rockets are summarized in table 2. Average particle speeds have been assumed for each sensor of each rocket until the distribution of orbits of dust particles can be determined. Most of the sensors on the rockets were possibly exposed to a high-speed-component of dust-particle influx; hence, higher particle speeds have been assumed for these sensors.

The direct measurements obtained with microphone systems on rockets, satellites, and spacecraft of the Soviet Union have been reported by Nazarova[15,16] and are summarized in table 3. Some of the quantities in table 3 have been computed on the basis of information given by Nazarova in order that data from space vehicles of both the United States and the Soviet Union can be included in the same analysis.

The sensitivities for the microphone systems on the Soviet spacecraft were expressed by Nazarova in terms of particle mass. The microphone system was assumed to be energy sensitive, and an average particle speed of 40 km/sec was assumed by her in converting to particle mass. An average particle speed of 40 km/sec was used in an early analysis by McCracken,[13] but this value is now regarded as being too high. A value of 30 km/sec seems more reasonable and will be used until information on the velocity distribution of the dust particles has been obtained. The mass sensitivities for the microphone systems on the Soviet spacecraft are therefore reduced by the square of 40/30 to compensate for the difference in the assumed particle speeds. The average particle speed assumed in computing the mass sensitivities of the microphone systems on the Soviet geophysical rockets was 15 km/sec, so the mass sensitivities given by Nazarova are increased by a factor of 4 when converting to 30 km/sec.

The influx rates measured by Sputnik III (1958δ2) underwen tremendous changes during the first 3 days of operation of the equipment. The influx rates, as reported by Nazarova,[15,17] were 4 to 11 particles/m^2-sec on May 15 (day of launch), 5×10^{-4} particles/m^2-sec on May 16 and 17, and less than 10^{-4} particles/m^2-sec during the interval May 18 to 26. Nazarova attributes the high-influx rates during the first few days to a meteoroid shower, but her conclusion is open to question. In any case, only the influx rate given for the last 9 days of operation can be used in establishing an average mass distribution curve. It is not clear whether or not Nazarova corrected the influx rate from Sputnik III for shielding by the earth, so the influx rate is left in table 3 as it was given.

The method of encoding information into the telemetered signal on Lunik I was such that only very crude upper limits to the influx rates can be specified. Only that influx rate measured by the scale of highest sensitivity is of any value in the present analysis. Lunik I (1959μ), Lunik II (1959ξ), and Lunik III (Interplanetary Station, 1959θ) operated at large geocentric distances, obviating corrections for shielding by the earth. No attempt has been made to correct the influx rates from the three geophysical rockets to omni-

TABLE 4.—*Direct Measurements from Photomultiplier Systems on United States Rockets and Satellites*

Vehicle	Half-angle of detector cone of vision (degrees)	Exposure (m^2-sec)	Number of particles	Influx rate (particles/m^2-sec)	Omnidirectional influx rate (particles/m^2-sec)
Aerobee NRL–25	80	0.63	101	160	390
Jupiter AM–28	80	2.2	4	1.63	4
Explorer VIII (preliminary)	60	4.3	110	25	200
Ranger I (preliminary)	75	8.5	179	21	114

directional values, because the orientations of the rockets and solid viewing angles of the sensors have not been reported.

DIRECT MEASUREMENTS FROM PHOTOMULTIPLIER AND ROCKET COLLECTION SYSTEMS

Experiments for measuring the kinetic energy of micron-size dust particles were flown on three rockets[9,18,19] and two satellites.[9,10] The sensors measured the intensity of the visible light emitted in each impact of a dust particle with a speed greater than 5 km/sec. Light emitted from impacts of microparticles with speeds between 4 and and 11 km/sec has been observed in laboratory experiments.[19] Measurements of the intensity and duration of the light flash provide a means for determining the kinetic energy of an impacting particle. Results of the laboraotry studies indicate that the light-flash sensor detected particles with masses greater than 10^{-13} gm.

The configuration of the sensors in each of the five experiments was different, but the principle of light-flash detection was the same. The basic detector unit was a photomultiplier tube. The surfaces exposed to impacts in the experiments were Lucite[18,19] and glass.[9,10] A few thousand angstroms of aluminum were evaporated on the impact surfaces to shield the photocathodes from background light. When a dust particle penetrated the aluminum during an impact, light from the impact flash could reach the photocathodes. The rocket experiments exposed larger impact areas than did the satellite experiments to compensate for the shorter exposure times of the rocket flights. The impact surface for each of the two satellite detectors was the face of an end-on type photomultiplier tube. An intensive effort was made to make the sensors insensitive to Cerenkov radiation and to energetic particles.

The results of the measurements with the light-flash detector are given in table 4. The cone of vision of the detector, the exposure, and the total number of impacts are given for each experiment. An impact rate is computed and then normalized to 4π steradians. (These computations also include correction factors for earth shielding.)

Aerobee NRL–25, launched at 0200 local time, exposed the light-flash detector to the high-speed component of the dust-particle influx; and Jupiter AM–28, at 1900 local time, to the low-speed component. The results may be normalized to 30 km/sec for purposes of comparison if the average particle speeds are taken as 45 and 12 km/sec, respectively. Since the detector sensitivity is a function of the square of the particle speed, the omnidirectional influx rates for the Aerobee and Jupiter—respectively 390 and 4 particles/m^2-sec (table 4)—become 173 and 25 particles/m^2-sec (assuming a linear relation between influx rate and mass sensitivity).

The Skylark rocket flown by Lovering[20] carried a light-flash experiment very similar to that on Aerobee NRL–25. No events were observed, but because of the following circumstances, there is a possibility that the experiment did not survive launch. The sensor was a 1P21 photomultiplier, and the rocket was a Skylark, which has a solid-propellant motor. The system was subjected to a 70-g acceleration test in a centrifuge. A meaningful environmental qualification test should have

also included severe shock and vibration tests, since the 1P21 is not ruggedized. An inflight sensor calibration with a light source would have verified the launch survival of the detector. The Jupiter AM–28 experiment and both satellite experiments contained this feature in the instrumentation.

The measurements from the satellites are corrected for earth shielding and normalized to 4π steradians. The data from the Explorer VIII and Ranger I experiments are restricted to satellite nighttime measurements in order to eliminate any extraneous counts caused by sunlight. The average omnidirectional influx rates for these measurements are 200 and 114 particles/m^2-sec, respectively, for particles with masses of 10^{-13} gm and greater.

Another estimate of the influx rates for dust particles somewhat smaller than those covered by the direct measurements obtained with microphone systems has been reported by Soberman, Hemenway, et al.[21] A recoverable high-altitude rocket was used in obtaining a very interesting collection of particles at altitudes greater than 88 km. An influx rate of 300 particles m^2-sec was estimated for particles with diameters equal to or greater than 3 μ. The cumulative influx rate (plotted as a function of particle diameter) has a negative slope of 1.3 and applies to particles with diameters as small as 0.2 μ. If a mass density of 3 gm/cm^3 is used,[21] the influx rates estimated from this collection may be compared to those obtained from the other direct measurements. No shielding corrections are introduced, since the collectors faced in the general direction of the apex of the earth's motion.

CUMULATIVE MASS DISTRIBUTION CURVE FROM DIRECT MEASUREMENTS

The direct measurements obtained with the microphone, photomultiplier, and rocket collection systems are plotted as a cumulative mass distribution curve in figure 2. Two characteristics of the data used in establishing the curve should be emphasized. First, the influx rates are expressed as omnididirectional values, and secondly, the curve is the result of a series of experiments in the vicinity of the earth. How well the curve applies to other regions of space is not presently known.

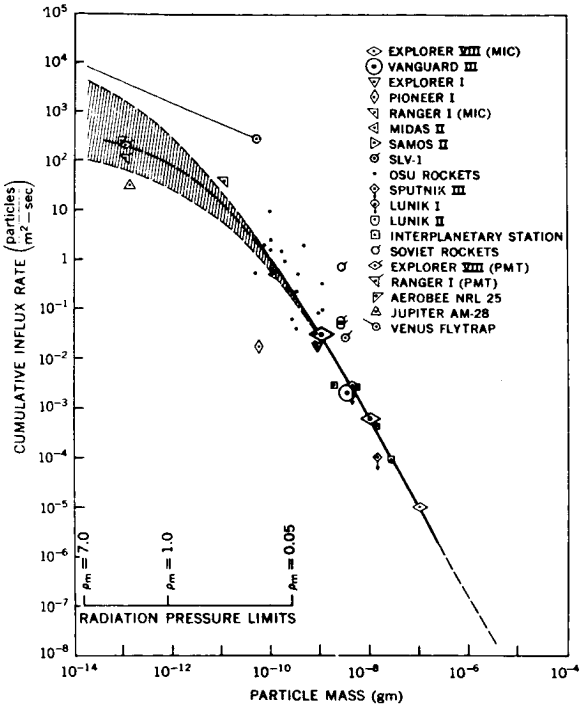


FIGURE 2.—An average cumulative mass distribution curve for the vicinity of earth derived from all the available direct measurements obtained with microphone and photomultiplier systems. The influx rates are omnidirectional values. A dust particle of mass density ρ_m is expelled from the solar system by radiation pressure if its mass is less than that shown on the lower abscissa opposite the particular value of ρ_m.

All of the microphone data for the mass range of 10^{-10} to 10^{-7} gm are compared in figure 2 with the Explorer VIII results (presented in figure 1). A study of the data points in figure 2 demonstrates the degree of consistency with which the average influx rates derived from all microphone measurements fit the curve.

The photomultiplier results allow an extension, to approximately 10^{-13} gm, of the distribution curve obtained with microphone systems. The reported influx rates from the rocket collection experiment[21] are somewhat higher than the photomultiplier detector results. The 80 to 150 km altitude range of the collection experiment probably contains an abundance of decelerated microparticles; hence, this is a possible explanation for the high influx rate of microparticles reported from the collection results.

The radiation pressure limits, in terms of particle mass density ρ_m, are also shown in figure 2

and subsequent figures above the particle mass abscissa. Each value of ρ_m determines a "cutoff" radius; all dust particles of this radius having masses smaller than that read below on the abscissa will be swept out of the solar system by radiation pressure. (The above statements are true only to the extent to which geometrical optics may be used in place of the more complex Mie theory of light scattering. The differences are not important in the present analysis.)

The cumulative mass distribution curve is not a constant mass-to-magnitude curve, and the slope appears to change rapidly with decreasing particle size. On a cumulative mass distribution curve, the slope should approach zero as the radiation pressure limits are reached. The results for particles with masses between 10^{-13} and 10^{-10} gm represent initial measurements which are inherently more uncertain than the microphone data. As the number of measurements increase for this range of particle size, the mechanisms controlling the distributions of these dust particles will be better understood.

RESULTS FROM PENETRATION AND FRACTURE EXPERIMENTS

Thirteen experiments with penetration or fracture types of dust particle sensors have been flown on seven U.S. satellites. In all but two cases, numbers of events detected by these sensors were one or zero; therefore, a measured average influx rate for dust particles cannot be determined. A comparison can, however, be made between the results of these experiments and the average mass distribution curve established by measurements with microphone systems.

The fracture type sensor consisted of continuous wire wrapped around an insulating support material. An impact was observed when a colliding dust particle fractured the wire and caused an open circuit. Manring[22] working with Explorer III (1958γ), Soberman and Della Lucca[11] working with Midas II and Samos II, and Secretan working with Explorer XIII (1961χ) have flown the wire grid type sensor on these five satellites. The major differences among these sensors were the wire size and the total exposed area. The mass of the dust particle which would fracture the wire is taken here as being comparable to that reported by Manring[22] and by Cohen et al.[23]

Four different types of penetration sensors have flown on three satellites. Three of the sensors required a perforation of the exposed surface, and one sensor required a crater with a diameter sufficient to destroy the sensing element. LaGow and Secretan[24,25] developed three of these sensors for Vanguard III.

The first type of sensor consisted of a chromium strip 300 μ wide and 1 to 3 μ thick, evaporated onto Pyrex glass. The resistance of the strip was monitored, and a complete break of the chromium was required to register an impact. The threshold sensitivity in terms of particle mass was determined by computing the diameter of the crater necessary to produce an open circuit.The second type of sensor consisted of two hermetically sealed and pressurized zones of which the exposed surface was 0.162 m^2 of the 26-mil thick magnesium skin of the satellite. A transducer constantly monitored the differential pressure between the two zones, so that a puncture of either or both could be detected. The third sensor consisted of a CdS cell covered by a 0.25-mil mylar film which was made opaque by evaporating aluminum onto both sides of the film. As penetrations occurred, the admitted sunlight changed the resistance of the CdS cell permitting measurement of the effective hole size. More than one penetration could have been observed with this sensor. This experiment was also flown on Explorer VII (1959 ι).

Davison[26,27] has flown the fourth type of penetration experiment on Explorer XIII. A plate of stainless steel was mounted in front of a foil gage consisting of a continuous strip of gold deposited on silicone rubber. The foil gage was separated from the metallic plate by a mylar insulator. Two thicknesses of stainless steel (75 and 150 μ) were used in two separate sensors. A particle sufficiently large to penetrate the metal plate and fracture the gold foil would have led to an open circuit and detection of an impact.

Using 0.25-mil mylar film and micron-size particles with speeds as high as 11 km/sec, Friichtenicht[20] has found that the diameter of the hole is 1.5 ± 0.5 times the diameter of the impacting particle for speeds greater than 3.5 km/

TABLE 5.—*Direct Measurements from Penetration and Fracture Experiments on U.S. Satellites*

Satellite	Type of dust particle sensor	Critical dimension for penetration or fracture	Range of threshold particle mass (gm)	Exposure (corrected) (m^2-sec)	Number of particles	Predicted influx rate $P_1=0.99$ (particles m^2-sec)	Date point shown in figure 3
Vanguard III	Pyrex-chromium strip	300 μ	4.0×10^{-8}–1.2×10^{-7}	7.0×10^{1}	0	6.2×10^{-2}	1
	Pyrex-chromium strip	300 μ	4.0×10^{-8}–1.2×10^{-7}	1.4×10^{2}	0	3.3×10^{-2}	2
	Mylar–CdS cell	0.25 mil	1.2×10^{-9}	8.7×10^{1}	0	5.3×10^{-2}	3
	Magnesium pressure zones	26 mil	3.0×10^{-7}–9.0×10^{-7}	7.2×10^{5}	0	6.4×10^{-6}	4
Explorer VII	Mylar–CdS cell	0.25 mil	1.2×10^{-9}	3.9×10^{1}	1	1.2×10^{-1}	5
Explorer XIII	Stanless stell–gold foil	75 μ	1.1×10^{-8}–3.3×10^{-8}	3.4×10^{4}	0	1.4×10^{-4}	6
	Stainless steel–gold foil	150 μ	8.8×10^{-8}–2.6×10^{-7}	8.5×10^{3}	0	5.6×10^{-4}	7
	Wire grids	75 μ	2.0×10^{-7}–6.0×10^{-7}	1.7×10^{3}	0	2.7×10^{-3}	8
	Wire grids	50 μ	6.7×10^{-8}–1.8×10^{-7}	7.7×10^{2}	0	6.0×10^{-3}	9
Explorer I	Wire grids	17 μ	4.2×10^{-9}–1.2×10^{-8}	3.6×10^{3}	0	1.3×10^{-3}	10
Explorer III	Wire grids	17 μ	4.2×10^{-9}–1.2×10^{-8}	2.4×10^{2}	2	6.4×10^{-3}	11
Midas II	Wire grids	20 μ	4.2×10^{-9}–1.5×10^{-8}	*2.0×10^{3}	0	2.3×10^{-3}	12
Samos II	Wire grids	20 μ	4.2×10^{-9}–1.5×10^{-8}	*1.1×10^{4}	8	8.0×10^{-4}	13

*Not corrected for earth shielding.

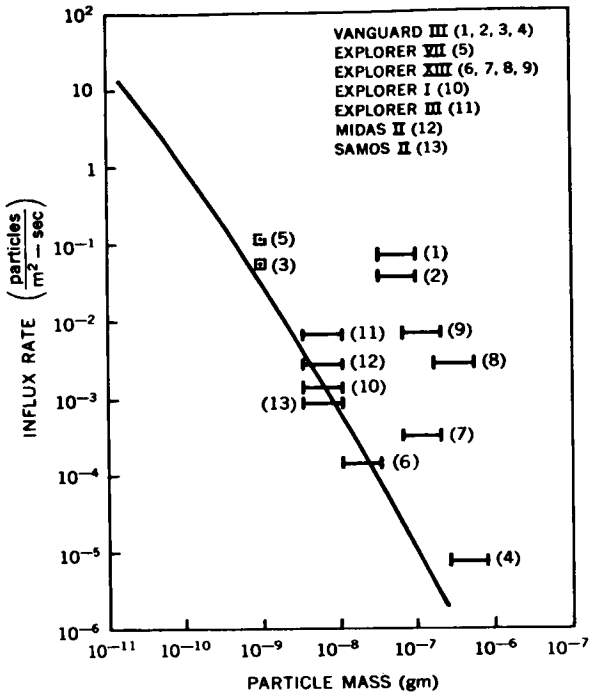


FIGURE 3.—Average influx rates computed (on the basis of a probability of 0.99 of at least one impact) from penetration and fracture experiments and compared with those given by the cumulative mass distribution curve. The numbers following the names of the vehicles identify the data points with the data given in table 5.

sec. Secretan and Berg (unreported results), using the same accelerator, have found no marked deviation from the above results. These results are the basis for the sensitivity used in the interpretation of the mylar-film CdS experiment. The other three penetration sensors (magnesium and stainless steel) required a perforation or a crater. An extensive series of penetration experiments has been performed by Summers et al.[28,29] The penetration equation developed empirically in these studies was used to compute the threshold mass sensitivity for the three sensors.

The pertinent information concerning the penetration and fracture experiments is summarized in table 5. The exposure (area-time product) includes a consideration of earth shielding, except for the two measurements by Soberman and Della Lucca.[11] The influx rates for Explorer III and Samos II were computed from the numbers of impacts and the corresponding exposures. One event was observed on Explorer VII, but no events were observed in the 10 remaining experiments. An average influx rate can be predicted for these latter experiments by computing the influx rate necessary for a probability of 0.99 for at least one impact. This is done in the following manner: an omnidirectional distribution of particles was assumed and the equation given below (based on Poisson statistics) was used to find the average influx rate:

$$P_1 = 1 - \exp(-atr)$$

where

$P_1 = 0.99$ = probability of at least one impact,

at = exposure (m^2-sec),

r = average influx rate (particles/m^2-sec).

The results derived from these experiments are compared in figure 3 with the average mass distribution curve established by measurements from microphone systems.

The indicated uncertainties represent a consideration of the major variations known at present concerning the parameters (particle speed and mass density) involved in hypervelocity impacts. The data show that some of the experiments had insufficient exposure to yield significant information. Within the uncertainties shown, the measurements do support the average mass distribution curve, especially within the mass range of 10^{-8} to 10^{-6} gm. The curve shown in figure 3 predicts the wire grid fractures which occurred on Explorer III[22] and Samos II[11] and the survival (without puncture) of the penetration experiments on Vanguard III[24] and Explorer XIII.[26]

The analysis of all of the direct measurements has shown no significant departures from the average mass distribution curve derived from the microphone experiment on Explorer VIII. Now that it has been demonstrated that the direct measurements are quite self-consistent, comparisons of the direct measurements with results obtained by other observational techniques can be attempted. Information deserving special attention in such a comparison comes from the observations of meteors, the zodiacal light, and the solar F corona.

COMPARISON OF DIRECT MEASUREMENTS WITH RESULTS FROM METEOR OBSERVATIONS

Extrapolating the results from meteor observations toward smaller particles represented a major method of estimating influx rates for dust particles before the advent of the direct measurements technique. The tabulation by Watson[30] of influx rates of meteors followed the constant-mass-per-magnitude relationship, rendering the extrapolation to smaller particles relatively easy. Such extrapolations have been made in the past, the most commonly used ones being those of Grimminger[31] and Whipple.[32]

Even the earliest of the rockets in the OSU series[13] gave influx rates several orders of magnitude higher than those expected on the basis of the extrapolations of meteor data. Meanwhile, an observation of coasting in the train of a double station photographic meteor led to the hypothesis of a very low value of mass density ($\rho_m = 0.05$ gm/cm^3) for meteoroids.[33] Whipple[34]—using this low value of mass density together with the corresponding change in the mass-to-magnitude relationship, the influx rates reported by Millman and Burland,[35] and an average speed of 28 km/sec for the meteoroids—proposed a constant-mass-per-magnitude extrapolation. This new distribution, showing much higher influx rates than are shown by the Watson distribution for particles of a given mass, did not agree with the early direct measurements.

The results of Millman and Burland[35] and of Hawkins and Upton[36] showed that the constant-mass-per-magnitude relationship was not valid, even for meteoroids. If the distribution curve given by Hawkins and Upton had been extrapolated into the direct measurements range of particle mass, it would have shown some degree of compatibility with the early results from the OSU rockets. Such an extrapolation was not in agreement with the direct measurements from the early satellites.[37,38]

On the basis of this early evidence of disagreement between the direct measurements and the extrapolated meteor results, McCracken (together with Alexander) first suggested that a constant-mass-per-magnitude law did not apply in the direct measurements range of particle mass.[13]

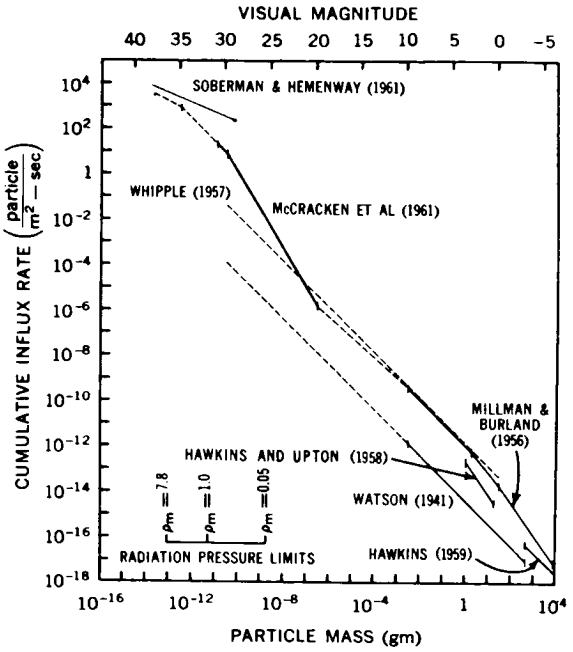


FIGURE 4.—Segmented cumulative mass distribution curve for omnidirectional influx rates of dust particles and meteoroids. The portions of the curve derived from the direct measurements of dust particles are known only to be valid for average conditions in the vicinity of the earth. The equations for the curve labeled "Millman and Burland" were given by McKinley (reference 39).

The direct measurements then available were not sufficient to confirm this hypothesis, nor did sufficient data become available until measurements were obtained with Explorer VIII. Actually, two interpretations of the direct measurements were possible at the time the suggestion was first made. They were: (1) The mass distribution curve obtained from the direct measurements departed significantly from those obtained by extrapolating results from meteor observations, or (2) the direct measurements were indicating the presence of a geocentric concentration of dust particles. Although the first interpretation seemed to be in better agreement with the early data, quantitative arguments for its validity were not possible until the data sample was obtained with the microphone system on Explorer VIII.

A segmented approximation to the cumulative mass distribution curve (fig. 2) is shown, together with several model distributions and observational

results from the meteor studies, in figure 4. The cumulative influx rates obtained by the various observational methods are plotted as a function of particle mass or visual magnitude, with the approximate radiation pressure limits for selected values of mass density being given as an auxiliary abscissa.

Visual magnitude has been introduced as an abscissa because the results from meteor observations can be placed on a mass distribution curve only within the limits set by the uncertainty ($\approx$200) in the mass-to-magnitude relationship. Visual magnitude is related to particle mass in figure 4 by assuming that the luminosity of meteors with a given speed depends linearly on the mass of the meteoroid and that a meteoroid with a mass of 25 gm and a speed of 28 km/sec will produce a meteor of zero visual magnitude.[34] Use of a mass density greater than the value of 0.05 gm/cm^3 used by Whipple[34] shifts the influx rate for meteoroids of a given mass toward a lower value.

The distribution curves for meteors given by Whipple,[34] McKinley,[39] and Hawkins and Upton[36] are plotted in figure 4 as functions of visual magnitude; a mass distribution given by Hawkins[40] for asteroids and fireballs is also included. The distribution given by Watson[30] is plotted in terms of particle mass. This is done so that the difference between the Watson and Whipple distributions can be used to illustrate the uncertainty in the influx rate of meteoroids of a given mass resulting from the poorly known mass-to-magnitude relationship for meteors. The direct measurements are not affected by this uncertainty. The uncertainties encountered in placing the direct measurements on a cumulative mass distribution curve are about two orders of magnitude smaller than those for meteoroids.

As figure 4 shows, the mass distribution curve obtained from the direct measurements departs markedly from those obtained by extrapolating results from meteor observations. It is evident that the constant-mass-per-magnitude relation does not hold for dust particles in the vicinity of the earth. There is, in fact, little evidence that the relation holds for more than a few magnitudes within any range of particle mass, except possibly for fireballs and asteroids.[40] Thus, there is no particular reason for trying to force the direct measurements to fit a constant-mass-per-magnitude curve.

The mass distribution of small dust particles was not known before the data were obtained with Explorer VIII. Several analyses (using best guesses, such as a constant-mass-per-magnitude relation) for the distribution curve in the direct measurements range of particle mass led to the conclusion that the direct measurements confirmed the existence of a geocentric concentration of dust particles. The results from Explorer VIII and the good agreement of all the other direct measurements with the Explorer VIII data show that these analyses are incorrect, because the measured mass distribution curve differs significantly from those assumed in the various analyses.

Some of the more subtle points shown in figure 4 can be illustrated more clearly by deriving from the incremental mass distribution the mass influx shown in figure 5. To remove the effects of using the segmented cumulative mass distribution (fig. 4) as a basis for deriving the incremental mass influx curve, the latter has been smoothed. It is given in terms of the accretion rate by the earth of dust particles or meteoroids in a given mass range as a function of particle mass or visual magnitude.

The most important conclusion which can be reached on the basis of the distribution curve shown in figure 5 is that the earth's accretion of interplanetary material is dominated by the small dust particles. The integrated accretion rate amounts to about 1×10^4 tons per day.

It must be realized that in both figure 4 and figure 5, the distribution curves for particles with masses less than about 10^{-11} gm are more uncertain than the segments of the curves derived from microphone results. Dust particles with masses less than about 10^{-9} gm are subject to the perturbative effects of radiation pressure. The shape of the distribution curves in the submicron range of particle size depends critically on the locations of the sources, on the distribution of orbits, and on the mass densities of the dust particles.

Gallagher and Eshleman[41] have found that the influx rates of faint radar meteors show large

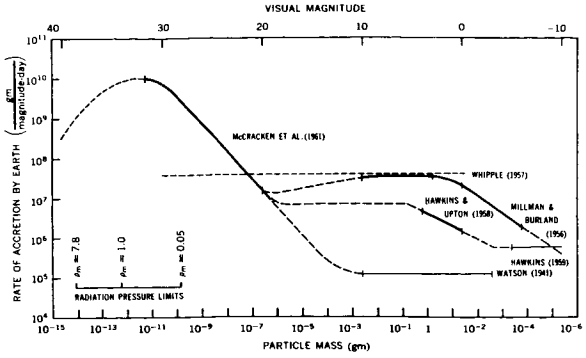


FIGURE 5.—Incremental mass influx curve derived from the cumulative mass distribution curve for interplanetary material accreted by the earth.

fluctuations with time. The observed grouping of radiants suggests that the faint radar meteors are members of "sporadic showers" rather than dispersed members of major meteor streams. Large fluctuations in the influx rate also seem to be the rule rather than the exception in the case of direct measurements. Dubin[7,8] has reported on the large fluctuations observed with Explorer I. Large fluctuations in the influx rate have also been reported for Vanguard III[42,43] and for Sputnik III,[15,17] although there is still some question about the latter case. (The fluctuations in influx rate observed on Vanguard III and Explorer VIII are presently under analysis.) It appears, on the basis of both the direct measurements and the data for faint radar meteors, that the dust particles are not nearly so uniformly distributed as are the sporadic meteoroids.

The interplanetary dust particle event detected by Explorer I on February 2 and 3 (shown in fig. 6) may be evidence of a "sporadic shower" of small dust particles, since it bears no relation to a known meteor shower. The large increases in influx rate which occurred during the November 16 to 18 interplanetary dust particle event observed with Vanguard III are shown in figure 7.

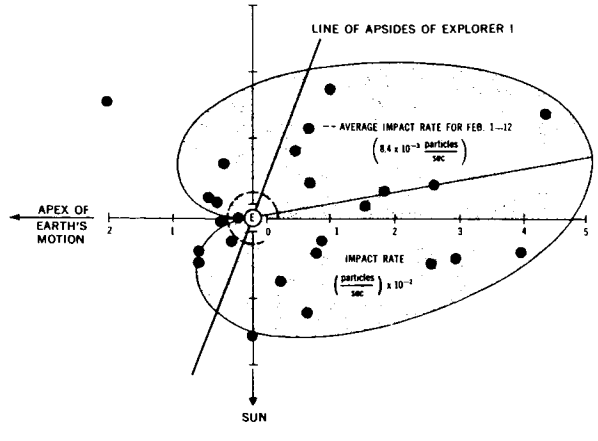


FIGURE 6.—Impact rates during the February 1958 interplanetary dust particle event plotted as a function of the position of the satellite relative to the earth.

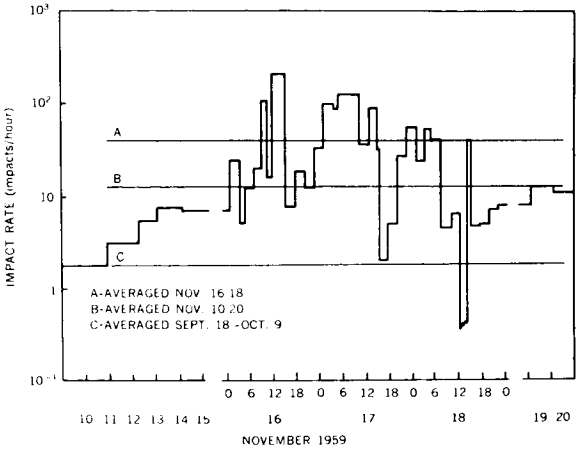


FIGURE 7.—Impact rates during the November 1959 interplanetary dust particle event. The impact rates during the three days of peak activity are shown on an expanded time scale.

The coincidence in time with the Leonids suggests that large numbers of small dust particles are being generated in the Leonid stream. The microphone system was almost omnidirectional, so it is not possible to establish the radiants of these dust particles. If they did belong to the Leonid stream, the imporbability of such dust particles remaining in the stream for one orbital period would require that they were released from larger meteoroids as they approached perihelion passage.

Eshleman[44] has reported that the influx rates of the faint radar meteors seem to increase rather than decrease as the limiting sensitivity of the equipment is approached. This may be evidence that the distribution curve obtained from the direct measurements can be extrapolated to join onto a distribution curve for meteoroids at about 15th visual magnitude (see fig. 4 or 5).

COMPARISON OF DIRECT MEASUREMENTS WITH RESULTS FROM ZODIACAL LIGHT AND SOLAR CORONA OBSERVATIONS

Photometric observations of the zodiacal light and the solar corona have yielded considerable information about the material in the zodiacal cloud surrounding the sun. The results obtained by analyzing the observations are generally expressed in terms of the spatial densities (and cariations in the spatial densities) of electrons and dust particles with distance from the sun. Invremental size distributions for the dust particles are also obtained in such analyses.

Direct measurements of the spatial density, mass distribution, and selected physical parameters of interplanetary dust particles have not yet been obtained for regions of space removed from the earth-moon system. The spatial densities and size distributions of dust particles inferred from photometric studies of the zodiacal light and solar corona presently represent the only available information about small dust particles in the zodiacal cloud.

Comparisons of the direct measurements with the results from zodiacal light and solar corona observations presently represent the only means of determining whether the available direct measurements obtained near the earth are also characteristic of interplanetary space. Dubin and McCracken[45] have compared the direct measurements with the results obtained in investigations of the zodiacal light and solar corona by van de Hulst,[46] Allen,[47] Elässser,[48] and Ingham.[49] It was found that if the results obtained by Ingham were taken as representative of interplanetary space, a spatial density near the earth at least three orders of magnitude higher than for interplanetary space must be introduced in order to remove the discrepancy. There is such a lack of agreement among the results from the various investigations of the photometric observations that comparisons of this kind are, at best, only qualitative.

CONCLUSIONS

The direct measurements obtained with the microphone system on Explorer VIII have provided a basis for analyzing all the available direct measurements of interplanetary dust particles. An average cumulative mass distribution curve, subject only to very minor uncertainties, has been established for dust particles in the vicinity of the earth. This average distribution is valid, within an order of magnitude or less, for particles with masses between 10^{-13} and 10^{-6} gm. The irregular shape of the distribution curve precludes the possibility of writing its equation in a simple analytical form.

The mass distribution curve obtained from the direct measurements differs markedly from those expected on the basis of extrapolations of results from meteor observations. Consequently, the earth's accretion of interplanetary matter may be said to be dominated by dust particles with masses less than about 10^{-6} gm. A conservative estimate of the accretion rate is 10^4 tons per day.

The influx rates obtained from the direct measurements undergo large fluctuations and, in one case, show a correlation in time with the expected annual return of a major meteor shower. These fluctuations suggest that the dust particles are not predominantly in long-lived orbits about the earth.

Discrepancies as large as 10^4 are found when the direct measurements in the vicinity of the earth are compared with the spatial densities of dust particles in interplanetary space inferred from photometric studies of the zodiacal light and the solar corona. However, such comparisons are uncertain because of the large discrepancies among the photometric data.

The available direct measurements are not sufficient to define either an average geocentric speed or an average mass density. These measurements encompass a range of particle mass which extends well past the radiation pressure limit for particles in heliocentric orbits with mass densities of 0.05 gm/cm^3. Mass densities of approximately 1 gm/cm^3 seem more reasonable for the direct measurements range of particle size. In view of the uncertainties concerning the probable origin (or origins) of the particles, the distribution of their orbits, and the mass densities of those of micron size, it does not seem wise to extrapolate results from the meteoroidal range of particle size to the smaller sizes of dust particles.

The various departures of the direct measure-

ments from expectations based on other methods of observation shows the feasbility of using the direct measurements technique to study material in the zodiacal cloud. Appropriate direct measurements will serve to answer most of the questions left unanswered in this analysis, and will represent an important means of determining the predominant source of the dust particles observed in the vicinity of the earth.

REFERENCES

1. Anderson, G. D., Burkdoll, F. B., et al., "Experimental Techniques Developed for Impact Studies of Microparticles," in: *Proc. 3d Sympos. on Hypervelocity Impact, Chicago, October* 1958, ed. by F. Genevese, Illinois Inst. Tech., Armour Res. Found., vol. 1, pp. 361–384, February 1959.
2. Stanyukovich, K. P., "Elements of the Theory of the Impact of Solid Bodies with High (Cosmic) Velocities," in: *Artificial Earth Satellites*, ed. by L. V. Kurnosova, New York: Plenum Press, vol. 4, pp. 292–333, 1961.
3. Lavrent'ev, M. A., "The Problem of Piercing at Cosmic Velocities," in: *Artificial Earth Satellites*, ed. by L. V. Kurnosova, New York: Plenum Press, vol. 3, pp. 85–91, 1961.
4. McCracken, C. W., Alexander, W. M., and Dubin, M., "Direct Measurement of Interplanetary Dust Particles in the Vicinity of Earth," NASA Technical Note D–1174, December 1961; also published in *Nature*, **192**(4801): 441–442, November 4, 1961.
5. ———, and ———, "The Distribution of Small Interplanetary Dust Particles in the Vicinity of Earth," NASA Technical Note D–1349, July 1962; also published in *Proc. Internat. Sympos. on the Astronomy and Physics of Meteors*, Cambridge, Mass, August–September 1961 *Smithsonian Contrib. Astrophys.* (to be published.)
6. Dubin, M., "IGY Micrometeorite Measurements," in: *Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January* 1960, ed. by H. K. Bijl, Amsterdam: North-Holland Publ. Co., pp. 1042–1058, 1960.
7. ———, "Meteoritic Dust Measured from Explorer I," *Planetary and Space Sci.*, **2**(2/3): 121–129, April 1960.
8. ———, "Remarks on the Article by A. R. Hibbs, 'The Distribution of Micrometeorites near the Earth'," *J. Geophys. Res.*, **66**(8): 2592–2594, August 1961.
9. ———, Alexander, W. M., and Berg, O. E., "Cosmic Dust Showers by Direct Measurements," in: *Proc. Internat. Sympos. on the Astronomy and Physics of Meteors*, Cambridge, Mass., August–September 1961 *Smithsonian Contrib. Astrophys.* (To be published.)
10. Alexander, W. M., and Berg, O. E., "Microparticle Hypervelocity Impacts from Ranger I." Paper presented at the 5th Hypervelocity Sympos., Colorado School of Mines, Golden, Colo., November 1961. (To be published.)
11. Soberman, R. K., and Della Lucca, L., "Micrometeorite Measurements From the Midas II Satellite (1960 ?1)," Air Force Cambridge Res. Labs., Geophys. Res. Directorate, GRD Res. Notes to. 72 (AFCRL 1053), November 1961.
12. LaGow, H. E., Schaefer, D. H., and Schaffert, J. C., "Micrometeorite Impact Measurements on a 20 in. Diameter Sphere at 700 km to 2500 km Altitude," in: *Annals of the International Geophysical Year*, London: Pergamon Press, vol. 12, part 2, pp. 465–472, 1961.
13. McCracken, C. W., "An Analysis of Rocket and Earth Satellite Measurements of Micrometeoric Influx," M.S. Thesis, Oklahoma State Univ., 1959; also Oklahoma State Univ., Res. Found., Final Rept. Contract AF 19(604)–1908, Appendix B, April 14, 1960.
14. Buck, R. F., "Acoustic Detection of Meteoric Particles," Oklahoma State Univ., Res. Found., Final Rept. Contract AF 19(604)–1908, April 14, 1960.
15. Nazarova, T. N., "The Results of Studies of Meteoric Dust by Means of Sputnik III and Space Rockets," in: *Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January* 1960, ed. by H. K. Bijl, Amsterdam: North-Holland Publ. Co., pp. 1059–1062, 1960.
16. ———, "Rocket and Satellite Meteoric Dust Investigations," *Proc. 12th Internat. Astronautical Congress, Washington*, 1961, ed. by R. M. L. Baker, Jr., and M. W. Makemson, Vienna: Springer-Verlag, 1962. (In press.)
17. ———, "Results of Exploring Meteoric Matter with Instrumentation of Sputnik III and Space Probes," in: *Proc. 11th Internat. Astronautical Congress, Stockholm*, 1960, ed. by C. W. P. Reutersward, Vienna: Springer-Verlag, vol. 1, p. 625 (Abstract), 1961.
18. Berg, O. E., and Meredith, L. H., "Meteorite Impacts to Altitude of 103 Kilometers," *J. Geophys. Res.*, **61**(4): 751–754, December 1956.
19. Friichtenicht, J. E., "Experiments with a Two-Million Volt Electrostatic Generator," Paper presented at the 5th Hypervelocity Sympos., Colorado School of Mines, Golden, Colo., November 1961. (To be published.)
20. Lovering, J. F., "Micrometeorite Impacts to an Altitude of 135 Km," *Planetary and Space Sci.*, **2**(1): 75–77, October 1959.
21. Soberman, R. K. (ed.), "Micrometeorite Collection From a Recoverable Sounding Rocket," Air Force Cambridge Res. Labs., Geophys. Res. Directorate, GRD Res. Notes No. 71 (AFCRL 1049), November 1961.

22. MANRING, E. R., "Micrometeorite Measurements from 1958 Alpha and Gamma Satellites," *Planetary and Space Sci.*, **1**(1): 27–31, January 1959.
23. COHEN, H. A., CORMAN, A., and DUBIN, M., "Calibration of Micrometeoritic Detectors Used in Satellites and Rockets," in: *Proc. 3d Sympos. on Hypervelocity Impact, Chicago, October* 1958, ed. by F. Genevese, Illinois Inst. Tech., Armour Res. Found., vol. 1, pp. 405–424, February 1959.
24. LAGOW, H. E., SECRETAN, L., and GIULIANI, J., "Experiments for Satellite Environmental Measurements," in: *Annals of the International Geophysical Year*, London: Pergamon Press, vol. 6, parts 1–5, pp. 319–321, 1958.
25. ———, and ———, "The Micrometeorite Penetration Experiment," in: *Juno II Summary Project Rept. Vol. I, Explorer VII Satellite*, NASA Technical Note D–608, pp. 263–272, July 1961.
26. DAVISON, E. H., and WINSLOW, P. C., Jr., "Direct Evaluation of Meteoroid Hazard," *Aerospace Eng.*, **21**(2): 24–33, February 1962.
27. ———, and ———, "Space Debris Hazard Evaluation," NASA Technical Note D–1105, December 1961.
28. SUMMERS, J. L., "Investigation of Hgih-Speed Impact: Region of Impact and Impact at Oblique Angles," NASA Technical Note D–94, October 1959.
29. NYSMITH, C. R., and SUMMERS, J. L., "Preliminary Investigation of Impact on Multiple-Sheet Structures and an Evaluation of the Meteoroid Hazard to Space Vehicles," NASA Technical Note D–1039, September 1961.
30. WATSON, F. G., "Between the Planets," Philadelphia: Blackiston, 1941 (rev. ed., Cambridge, Mass.: Harvard University Press, 1956).
31. GRIMMINGER, G., "Probability That a Meteorite Will Hit or Penetrate a Body Situated in the Vicinity of Earth," *J. Appl. Phys.*, **19**(10): 947–956, October 1948.
32. WHIPPLE, F. L., "Meteorite Phenomena and Meteorites," in: *Physics and Medicine of the Upper Atmosphere*, ed. by C. S. White and O. O. Benson, Jr., Albuquerque: University of New Mexico, pp. 137–170, 1952.
33. ———, "Some Problems of Meteor Astronomy," in: *Radio Astronomy*, ed. by H. C. van de Hulst, Cambridge: The University Press, pp. 375–389, 1957.
34. ———, "The Meteoritic Risk to Space Vehicles," in: *Proc. 8th Internat. Astronautical Congress, Barcelona*, 1957, ed. by F. Hecht, Vienna: Springer-Verlag, pp. 418–428, 1958.
35. MILLMAN, P. M., and BURLAND, M. S., "The Magnitude Distribution of Visual Meteors," Paper presented at the 96th meeting of the American Astronomical Society, New York, December 1956.
36. HAWKINS, G. S., and UPTON, E. K. L., "The Influx Rate of Meteors in the Earth's Atmosphere," *Astrophys. J.*, **128**(3): 727–735, November 1958.
37. DUBIN, M., "Direct Measurement of Meteoric Dust Using Rockets and Satellites," in: *Trans. 10th General Assembly Internat. Astronom. Union, Moscow, August* 1958, ed. by D. H. Sadler, Cambridge: The University Press, pp. 714–715, 1960.
38. NAZAROVA, T. N., "Rocket and Satellite Investigation of Meteors," in: *Trans. 10th Assembly Internat. Astronom. Union, Moscow, August* 1958, ed. by D. H. Sadler, Cambridge: The University Press, pp. 713–714, 1960.
39. MCKINLEY, D. W. R., *Meteor Science and Engineering*, New York: McGraw-Hill, 1961.
40. HAWKINS, G. S., "The Relation Between Asteroids, Fireballs, and Meteorites," *Astronom. J.*, **64**(10): 450–454, December 1959.
41. GALLAGHER, P. B., and ESHLEMAN, V. R., "'Sporadic Shower' Properties of Very Small Meteors," *J. Geophys. Res.*, **65**(6): 1846, 1847, June 1960.
42. LAGOW, H. E., and ALEXANDER, W. M., "Recent Direct Measurements of Cosmic Dust in the Vicinity of the Earth Using Satellites," in: *Proc. 1st Internat. Space Sci. Sympos., Nice, January* 1960, ed. by H. K. Bijl, Amsterdam: North-Holland Publ. Co., pp. 1033–1041.
43. ALEXANDER, W. M., MCCRACKEN, C. W., and LAGOW, H. E., "Interplanetary Dust Particles of Micron-Size Probably Associated with the Leonid Meteor Stream," NASA Technical Note D–1154, December 1961; also published in *J. Geophys. Res.*, **66**(11): 3970–3973, November 1961.
44. ESHELMAN, V. R., and GALLAGHER, P. B., "Radar Studies of 15th-Magnitude Meteors," *Astronom. J.*, **67**(5): 245–248, June 1962.
45. DUBIN, M., and MCCRACKEN, C. W., "Measurements of Distributions of Interplanetary Dust," *Astronom. J.*, **67**(5): 248–256, June 1962.
46. VAN DE HULST, H. C., "Zodiacal Light in the Solar Corona," *Astrophys. J.*, **105**(3): 471–488, May 1947.
47. ALLEN, C. W., "The Spectrum of the Corona at the Eclipse of 1940 October 1," *Monthly Not. Roy Astronom. Soc.*, **106**(2): 137–150, 1946.
48. ELSASSER, H., "Fraunhoferkorona and Zodiakallicht," *Zeitschrift fur Astrophysik*, **37**(2): 114–124, July 13, 1955.
49. INGHAM, M. F., "Observations of the Zodiacal Light From a Very High Altitude Station. IV. The Nature and Distribution of the Interplanetary Dust," *Monthly Not. Roy. Astronom. Soc.*, **122**(2): 157–176, 1961." (See also the three preceding papers by D. E. Blackwell and M. F. Ingham, pp. 113–155.)

THE PLASMA IN INTERPLANETARY SPACE*

L. BIERMANN†

Max Planck Institute for Astrophysics

Interplanetary plasma, including the various lines of evidence for it, is described. This is followed by an account of the structure and composition of comets with emphasis on the ionized tails. The characteristics of these tails are explained by their interaction with a "solar wind" type plasma. Cometary phenomena bearing on the spatial and temporal distribution and dynamics of the plasma are discussed. A space experiment consisting of the formation of an artificial plasma cloud whereby these ideas could be tested and the plasma further investigated is described. Calculations indicate that the release of about 100 g of barium at 200,000 km could be observed.

INTRODUCTION

The subject of plasma in interplanetary space has been attacked by astronomical observation and theoretical work for a number of years. It is gratifying therefore that experimental research on it has now become possible by means of space probes and satellites with distant apogees.

Interplanetary space can be defined literally as the space between the sun and the planets. There has been some discussion in recent years as to what extent it should be regarded as an extension of the solar corona. We shall not, at present, go into the problem of how to define the boundary between the solar corona and interplanetary space. Let us just take the position that the corona, as is suggested by its appearance during a total solar eclipse, stretches outward to about 10 and possibly 20 solar radii, and that beyond this distance we reach interplanetary space.

The question of the physical distinction between the solar corona and interplanetary space is rather difficult. Leaving aside the radial motion, there is one basic difference as far as the co-rotation with the sun is concerned. The visible corona *does* co-rotate with the sun; this means that it rotates once in every 25 days. From the work to be discussed here, it is evident that interplanetary material, at distances of the order of 1 AU, does not co-rotate. One of the tasks before us, then, is to determine in detail the state of rotation of the material found in interplanetary space, as a function of position and time. This question cannot as yet be answered from observation.

How is plasma in interplanetary space observed? In addition to the indirect evidence from magnetic storms (see below) and the recent direct measurements there is *first* the body of information from the *comet tails*. Comets will be discussed later, in more detail; let us just say now that there is one type of tail, which is—by convention—called a Type I tail. These tails are essentially straight and long, directed approximately radially away from the sun, and are found to consist of ions of CO and N_2 and of some other ionized molecules in overall densities ranging from a few to perhaps 10^2 or 13^3 particles/cm^3; however there are usually structures in which the density is considerably higher; and the density also varies with distance from the head. From plasma physics we know that this material must necessarily constitute a quasi-neutral gas; that is, there must be as many electrons per cubic centimeter as there are ions.

*Presented at a NASA Goddard Space Flight Center Colloquium on October 17, 1961, Greenbelt, Maryland. This talk together with the question-answer period which followed (Appendix A) is published here in approximately the form of the original presentation; for this reason references to other work in the field are presented as originally given.

†Dr. Bierman is Director of the Max Planck Institute for Astrophysics, Munich, Germany.

These molecular ions move away from the sun at velocities of the order of 30 up to several hundred kilometers per second. This is the kind of plasma that for several decades has been known from observation to exist in interplanetary space in the place where there is a comet with a Type I tail.

For a large comet, the statistical probability of having a Type I tail is something like 50 percent; this estimate is based on evidence taken from a total of more than 50 comets [This is mostly from the work of Bredichin-Jaegermann (Petersburg 1903) as discussed by P. Stumpff, *Astr. Nachr.* 286: 87(1960). A new discussion on the basis of a very much larger material has been started.]

Second, there are the direct measurements of the *interplanetary plasma* from satellites or space probes; the most significant work in this area has been done recently by Bridge, Rossi, and their coworkers at the Massachusetts Institute of Technology. Some important work has been done also in Russia; this, in a general way, confirms the MIT results but doesn't give as many details.

At this stage it is sufficient to say that, during geomagnetically quiet periods, in the vicinity of the earth (at a distance of approximately 200,000 km) there are densities of the order of 10 ions (and electrons) per cubic centimeter; these ions move radially away from the sun with velocities on the order of 300 km/sec.

Third, we have evidence from the observation of *radio sources* near the sun. If by the apparent motion of the sun along the ecliptic during the year a radio source happens to lie on a line passing near the sun, say within 50 or 100 solar radii (equivalent to 12 to 25 degrees of arc) from the sun, it has been observed that the apparent size of the radio source is affected in a way that indicates the influence of some sort of scattering.

A more detailed discussion shows that there must be scattering by electrons in interplanetary space. While nothing can be derived about their state of motion, some rather indirect information may be gained about the density distribution.

Fourth, the observation of *zodiacal light* yields some evidence. Quite a number of investigations on this subject have been reported in the literature; but much of the earlier interpretation of the data has been superseded by later research.

The only reasonably certain conclusion at the present time is that during magnetic storms there might be electron densities of the order of some 10^2 electrons/cm^3.

There is also the general evidence from geomagnetism. Chapman and Ferraro have shown how magnetic storms are caused by plasma streams emitted from the sun and moving through interplanetary space with a velocity on the order of 1000 km/sec. But evidence as to the densities of these plasma streams is rather meager, and it is very difficult to give more than a lower limit.

Also, from geomagnetic evidence, two types of events should be distinguished. First, the magnetic storms: These are individuals events often connected with chromospheric eruptions on the sun. Second, there are recurrent streams—usually phenomena of a smaller scale—that have a tendency to repeat themselves after 27 days, possibly three times, four times, or even more often. These are presumably due to persistent regions on the sun called M-regions, which are believed to be semi-permanent sources of rather intense corpuscular emissions, their lifetime being on the order of several months or more.

The question of whether—in addition to the individual big events and the recurrent but quasi-stationary emission—there is also a stationary, more or less permanent and roughly isotropic outflow of coronal gas cannot be answered positively from geomagnetism. There has been a growing tendency in geomagnetism to connect the fluctuations of the geomagnetic field at polar stations, which are practically always observed, with matter coming from the sun. It may be expected that the observations made during the IGY will settle this question.

The main question that we will discuss today is this: Excepting the Type I tails of comets, the isolated solar eruptions that cause magnetic storms, and the recurrent streams of solar matter, is there always some plasma moving through interplanetary space? What we hope to show is that the observations of the comets' tails do indeed give a rather strong reason to believe that there is practically always some outflow of ionized matter from the sun.

COMETS AND COMET TAILS

The evidence from the comets and the interpretation of the acceleration and ionization of the cometary material is now given somewhat more in detail. Also, some plans for using the processes in the comet's tails as a model for an experiment will be discussed briefly. For this experiment, artificial plasma clouds are to be produced in interplanetary space. These should be directly observable and should behave in a way similar to the comets' plasma tails—and serve as probes for the presence of interplanetary material and its state of motion.

As a background to this discussion, a brief description of a comet will be helpful. First, a comet is believed to have a solid nucleus with a diameter, say, of a few to 10 or 20 kilometers, which consists of dust and of molecules of C, N, O, and H in frozen form. There is a good reason to assume that all this material gives rise to the appearance of the comet during its near approach to the sun, when this more or less solid nucleus is under the influence of solar light and heat. The nucleus is not very solid in the sense that the molecules present probably form a sort of ice that is frozen together with the dust; this model of a cometary nucleus has been developed and discussed particularly by Whipple at the Smithsonian Institute.

If a comet approaches the sun to within a few astronomical units, it gets a coma consisting of nonionized gaseous molecules like CN, C_2, C_3, CH, NH, and OH, and in addition one or two tails. In the latter there are dust particles and/or ionized molecules like CO^+ and N_2^+. The connection between composition and the type of the tail is this: The long, straight Type I tails are composed of ions, and other more curved types of tails are composed of dust and possibly some nonionized molecules. But here we shall discuss mainly the Type I tails.

Around the solid nucleus of the comet is the coma, very roughly 100,000 to a million kilometers in diameter. Furthermore, we may have a straight, extended tail that would be of the Type I and/or a shorter, curved tail that would be of Type II or III. These latter two types need not be distinguished for our purpose, but it should be noted that their direction with respect to the sun is usually different from that of the Type I tail. Type I tails are the longest, their length being of the order of about 10 million kilometers, and in some cases even more.

Both the dust and the nonionized molecules are subject only to a moderate or small acceleration away from the sun, whereas the ionized particles are driven away from the sun much faster. The former show accelerations away from the sun that are comparable only to the acceleration of gravity (0.6 cm/sec^2 at 1 AU). In contrast, the acceleration acting on the constituents of Type I tails—that is to say on the ions—is very much larger, say by a factor between 20 and 100. In individual cases it appears to go up to 1000 cm/sec^2, and even more.

This general picture explains the difference of the shapes. The motion of particles under the influence of solar gravitation is essentially a problem of celestial mechanics that was solved by astronomers during the last century. Thus one could explain, in a general way, the shapes of the tails of the Type II and III—and in less detail also the shape of the Type I tails.

With regard to the mechanism of acceleration away from the sun, it should be added that the pressure of the ordinary solar light is quite sufficient to account for the observed features of the tails of Type II or III. For dust particles that are small enough, the acceleration by light pressure is of the right order of magnitude; for nonionized molecules we may readily compute the light pressure from the transition probabilities in question, and again the acceleration turns out to be on the order of unity. Hence there is no particular problem in accounting for the observed acceleration of the nonionized molecules and of the dust particles.

There is, however, a serious problem concerning the ionized particles, because the transition probabilities of the resonance transitions in question in this case are quite small. The best value for the oscillator strength of these transitions is only 0.002. That means that the light pressure on these ions can account only for something like 0.1 cm/sec^2. So, if the light pressure were operative, we would expect an acceleration that would be smaller by about one power of 10 as compared

with the CN or C_2. What we observe *in fact* is that it is very much larger.

First, some of the main properties of these Type I tails should be mentioned. They are approximately parallel to the radius vector from the sun, as was stated before. In the second approximation there is a phenomenon of quite some interest: It was found first by Hoffmeister that the main axis of the comet deviates slightly from the radius vector so that the tail is lagging behind, in the sense of the comet's motion around the sun, by a few degrees of arc—say, by something like 3 to 5 or 6 degrees. This is significant from the theoretical point of view, which we shall develop later.

Next we have the extent of these tails, mentioned already. While the diameter of the tail is on the order of some 10^5 to 10^6 km, which is comparable to the diameter of the coma, the length may be very much larger. There have been a few comets possessing tails that extended even through a large part of the solar system, up to distances comparable with the distance between the earth and the sun.

There are usually structures in the tails of Type I. For example, there are elements that might be described as clouds and that can easily be followed on successive plates from one hour to the next—and sometimes from one night to the next, or even longer.

Velocities can be derived by measuring structures, e.g., "clouds" appearing on successive photographic plates. The values are of the order of 20 or 30 to 50 km/sec, but sometimes are several hundred to 100 km/sec. The ordinary range, however, is 30 or 50 to 200 km/sec.

An average value of the acceleration can also be found by plotting the velocity as a function of the distance. The time scale of the motion of an individual element (that is, the time during which we may say that a given element of mass is visible in the comet itself) is of the order of some days. This means, furthermore, that the tail is a transient phenomenon; for while the time of visibility of every comet is of the order of, say, several months—depending on the perihelion distance and the other conditions, the material that can be seen one night in the tail is replaced for the most part after a week, and certainly completely so after a fortnight.

Another interesting and frequent phenomenon is presented by very narrow streamers, a regular feature of many tails. These are structures with a diameter of only a few thousand kilometers and a length of the order of 10^6 km. The fact that they are so narrow is again significant: If the material would expand sideways with the velocity of sound, then a very low temperature of the order (of a few degrees Kelvin) would be required to keep it that narrow for the length of time given by the extent and the velocity. So there must be some lateral force that confines the streamers; and the only likely explanation is that there are magnetic fields associated with the streamers which, by their pressure, keep the plasma together.

A particularly important feature of these tails is what might be termed their activity. If several successive pictures of a Type I tail are compared, the variations from one plate to the next are usually quite conspicuous: Part of the material is moving farther away from the coma, the shape is changing, and material that has been at some distance from the axis of the tail has often come nearer to it.

If we compare pictures taken during successive nights, then we see that the whole appearance of the tail usually has changed considerably. Part of the tail material has been largely replenished, and at the same time there has been considerable displacement and rearrangement. A structure seen on one side of the tail axis during one night may even have moved to the opposite side of the axis on the next night; there may be in such cases, in addition to the displacement along the axis, a rotational motion around it. The plates of the comet Mrkos give an example of this phenomenon.

MECHANISMS OF PLASMA–TAIL INTERACTION

Briefly stated, the main point of the hypothesis that was introduced just 10 years ago is as follows.

The acceleration of the plasma, which forms the Type I tails, and the "activity" of these tails are attributed to the influence of solar particle radiation. The large acceleration is ascribed to the fact that, from plasma physics, the interaction

between two interpenetrating streams of plasma is known to be large, especially in the presence of magnetic fields. The tail plasma is moving, at least initially, quite slowly as compared with the particle streams coming from the sun. We would expect on quite general grounds that there will be a coupling between these two plasmas.

One basic consideration is that, if we ascribe the acceleration of these tails to the influence of solar corpuscular radiation, then their activity is a consequence of solar activity. Ordinary solar light is constant within a fraction of 1 percent. So we would not expect any variation of the acceleration because of the pressure of solar light acting on the structures composed of nonionized molecules and of dust.

On the other hand we know from geophysics that the particle radiation of the sun shows a great deal of variation, even between magnetic storms, and at a generally low level of solar activity; so it is natural to assume that the activity of the comet's plasma tails just reflects this type of activity of the sun—the activity which shows up on the earth as magnetic activity and in various other geophysical phenomena.

Now, in order to give substance to this hypothesis, let us first check whether the momentum flow coming from the sun is really sufficient. Instead of the line of reasoning that was current some years ago, the evidence from the measurements of Bridge, which give perhaps a lower limit to the particle flux, will be used. If we take the proton density and the velocity that Bridge measured—the latter on the order of several hundred kilometers per second, we get a momentum flow from the sun under quiet solar conditions of 10^{-8} dyne/cm^2.

Again, if we take average figures for a Type I tail at some distance from the head and the lower velocities observed in these tails under quiet conditions, we get values on the order of some 10^{-9} dyne/cm^2.

We know that the momentum flow from the sun is very much increased in the recurrent streams observed during low levels of solar activity or during magnetic storms. It is somewhat difficult to write down a precise figure, but all evidence points to increases in the momentum flux of up to several powers of 10. These, of course, would show up in the comet's tail by more violent phenomena.

The momentum flow, then, seems adequate to accelerate the comet's tail provided the coupling between the tail plasma and the solar plasma is sufficiently strong.

With regard to the mechanism of coupling, there are three possibilities. One is that coupling is due merely to the thermal motion of the electrons; this scheme was discussed in detail in 1951. It can be shown to be operative if the density of the solar stream is on the order of 10^2 or 10^3/cm^3 and if the electron temperature is on the order of 10,000°K. While in 1951 there was some observational evidence pointing toward densities in this range, we believe now that these are not there under average conditions.

Second, there is the possibility also discussed 10 years ago, and rediscussed by several authors since then, that magnetic fields couple together the solar plasma and the plasma of the comet's tail. This possibility appeared fairly remote in 1951, but in the meantime we have much indirect evidence about the presence of magnetic fields in interplanetary space; furthermore there are the direct measurements by means of space probes or satellites, such as those by Dr. Heppner. In addition, there is the evidence from the very fine streamers which, as was noted, give strong reason for believing that there are often (perhaps regularly) magnetic fields in the Type I tails. We may ask, then, whether the magnetic stresses are large enough to account for the observed transfer of momentum.

Since the magnetic fields that have been observed directly under quiet conditions are of the order of 10 to 20 gamma (1 to 2×10^{-4} gauss), evidently there are magnetic stresses on the order of 10^{-9}, or several times 10^{-9} dyne/cm^2; this seems adequate to account for the observed momentum flux.

Third, it might be that specific plasma effects due to plasma instabilities, such as fluctuations of the electric fields, are involved here. This is a possibility that has become apparent recently in connection with the experiments in plasma physics related to the problem of controlled fusion.

The state of our knowledge in this area is not yet sufficient to allow a definite statement as to whether or not this type of plasma coupling should effectively increase the coupling between a solar particle stream and cometary plasma.

OBSERVED PHENOMENA

Let us now consider observational effects that might, or might not, support this picture.

First of all, of course we have to ask whether we find correlations or recurrence phenomena (due to the sun's rotation) similar to those known in geomagnetism. If, for instance, we have a situation in which the earth and the comet are (in this order) approximately along one radius vector from the sun, then we would expect that, whenever we observe a really big magnetic storm on the earth, we should also observe something spectacular happening in the comet's tail.

That is to say, we should expect a close correlation between geomagnetic phenomena and the phenomena in the comet's plasma tail. Of course, the condition described above does not happen often. But there have been some such cases—for instance, the comet 1942g, Whipple–Fedtke; the comet was moving approximately in the plane of the ecliptic, and furthermore the perihelion of the comet coincided approximately with the opposition to the sun as seen from the earth. This is another favorable geometrical condition since, when the comet is in perihelion, the earth might be in any longitude; so it is really by chance that they both are in approximately the same heliocentric longitude. As a consequence, the comet could be seen in approximately the same direction of the sky (relative to the sun) for a fairly long time.

This was a time which there were only a few magnetic storms. There was one particular storm, the largest for about one-half year, on March 29, 1943, which showed up in the comet's tail by an exceptionally large acceleration.

Good observations of this comet were obtained at the Sonneberg Observatory. Values of the acceleration were derived for a number of individual days by the method that we described before.

There are two dates in 1943 where the motions in the tail were exceptionally fast and turbulent—March 3 and March 29. (The comet had been found late in 1942, but observations were mainly obtained during the first months in 1943.) Now it can be seen that the interval between these two dates is approximately one period of the sun's rotation.

Here the synodic period of rotation of the sun depends on the motion of the body to which it is referred. The sidereal period—that is, the true period of rotation of the sun—is 25 days. But, since the earth is moving in the same sense as the sun rotates, the sun appears to rotate only with a period of 27 days. For the comet Whipple–Fedtke, it would have been 26.5 days. Thus, this is the recurrence period to be expected if the high degree of turbulence observed on March 3 and 29 was due to the corpuscular radiation emitted by the same active region on the sun. As has been noted, on one of these two dates (March 29) there was a major magnetic storm on the earth that showed up very drastically in the structure of the tail. In contrast, on March 3 the event in the tail of the comet was not accompanied by a particular event in the geomagnetic records but there was another geomagnetic event 27 days after March 29. Of course, there was always some angle (20 to 30 degrees) between the comet and the earth, as seen from the sun; it is known that the streams coming from the sun have only an angle of opening of the same order.

In consequence, the observed correlation with the geomagnetic storm of March 29th and the apparent recurrence period of 26 days is probably just what we should expect under these circumstances. [See the pictures reproduced in *Z. f. Astrophys.* 54: 67 (1962) and 22: 286 (1943), especially pp. 304, 305.]

Another investigation was made for the comet Halley as observed during its reappearance in the year 1910. Again it was possible in observing the comet to find a synodic period of rotation of the sun as seen from the comet Halley. In that case it was only 22 days, because the comet Halley moves around the sun in a direction opposite that of the earth; therefore, the value is less than 25 days. Also, it was possible, by an appropriate analysis, to get some correlation with geomagnetic phenomena; however, this was somewhat more difficult because the general activity of the sun was at a higher level than 1943. Suffice

it to say that the general result was again in agreement with that anticipated.

There are a few more events of this type that have been discussed, mainly by Rhea Lüst in some recent papers. The expected correlation between cometary and geomagnetic events and recurrence phenomena in analogy to those known in geomagnetism thus have been found in the plasma tails of the comets.

Furthermore, as mentioned before, the tail in general appears to lag behind the radius vector from the sun by a small angle, in the sense of the comet's orbital motion. That, again, is precisely what we would expect in case of a mechanical transfer of momentum if the material is coming from the sun with a speed of, say, approximately 10 to 20 times the orbital velocity of the comet—that is, a few hundred kilometers per second.

If instead we have electromagnetic coupling, the situation is somewhat more complicated; but from the momentum balance we would still expect an effect of this general character. Se we may take the observation of this angle as another qualitative verification of the basic idea.

If, then, we regard the hypothesis that the acceleration of the plasma tails of comets is due to the interaction with the solar corpuscular radiation as verified, we can immediately draw one or two conclusions not obvious from geomagnetism. The first is connected with the fact that if a comet is once observed to have a plasma tail, then it is found to possess such a tail, though with varying properties, subsequently during the whole time of its near-approach to the sun. The tail is ordinarily visible as long as the distance from the sun is smaller than about $\frac{3}{2}$ AU. That is to say, if a comet gets this type of tail at all, it tends to have it for the whole time it is inside (approximately) the orbit of Mars.

But, as we have said, of the brighter comets only one in approximately two does get an observable plasma (Type I) tail; presumably the property of getting such a tail depends on the chemical composition and possibly also on the past history of the comet.

It should perhaps be mentioned here that a given comet may make something like 100 near-approaches to the sun. There are about 10^{11} potential comets in the sky, most of which are moving at distances almost comparable with the distances of the nearest stars. All these bodies move through intersteller space with the sun and are hence actually members of the solar system. But, through the gravitational disturbances caused by stars passing nearby, their constants of motion are changed statistically; and so it happens that one or two every year are thrown into orbits that touch the inner part of the planetary system. As soon as they approach the orbit of Jupiter or Saturn, their constants of motion are changed in such a way that they either disappear entirely (by getting into an hyperbolic orbit around the sun) or, with equal probability, they get into an orbit having a major axis of the order of 20 or 30 AU (e.g., comet Halley).

From then onward the comet may be in this orbit for about 50, 100, or 200 revolutions around the sun. But the nucleus, losing something like 1 percent of its material in each approach to the sun, is not a permanent body. In a few cases we have witnessed the breaking up of a comet during its near-approach to the sun.

It has been mentioned that a comet with a Type I tail appears to have this tail during most of the time spent in the inner part of the planetary system. That means, then, that the solar corpuscular radiation is not there exclusively during the time of magnetic storms or recurrent magnetic phenomena but that there must be some corpuscular radiation all the time; otherwise, we would have to assume a quite unlikely correlation between the appearance of comets and the production of solar corpuscular radiation. For this component of the solar corpuscular radiation, Eugene Parker has proposed the term "solar wind"; this seems a very appropriate term as long as it is not used also for the more violent phenomena of recurrent character or of the magnetic storm type.

We observe furthermore that comets with Type I tails are not confined to periods of high solar activity. They are actually rather scattered through the whole 11-year cycle, although there may be some dependence on phase in this cycle as a second-order effect. Out of the total number of something like 30 comets with Type I tails

which have been looked at more closely so far, it is not possible to give a more precise estimate.*

The comets are moving around the sun not only in the neighborhood of the plane of the ecliptic, but in orbits that may have any inclination to that plane: There have been comets that passed the vicinity of the poles. We might therefore wonder whether there is any difference in the properties of these comets as compared with those moving near the ecliptic plane.

A preliminary study of this kind has been made by Peter Stumpff [*Astron. Nachrichten* 286: 87 (1962)]. He separated the comets for which enough data were available into several groups according to the heliographic position of their perihelion.

Dividing the sky into two areas of equal angular area—one centered around the ecliptic plane (approximately the equatorial plane of the sun) and the other half around the poles, Stumpff found that the ratio of comets with plasma tails to the comets without such tails differed by a factor of 2. The over-all probability of a comet having a plasma tail seems to be something like $\frac{2}{3}$ if its perihelion is within 30 degrees of the ecliptic plane as compared to about $\frac{1}{3}$ if its perihelion is within 60 degrees of the poles.

This result appears to be consistent with evidence from the occultation of radio sources, which again gives the impression that the scattering electrons in interplanetary space have nonspherical surfaces of equal density around the sun; these observation indicate some flattening of these surfaces out to 60 to 100 solar radii, but not at all an absence of scattering electrons at high latitudes. The electrons in question should, of course, belong to the solar corpuscular radiation.

There is another conclusion: We observe that the momentum transfer from the solar particle streams to the plasma of the comet is to the first order of approximation along the radius vector from the sun; to the second order we have definite indication that the individual motion of the comet plays some role. This means that velocities on the order of 20 or 30 km/sec perpendicular to the radius vector really do show up in the general direction of the tail.

*See also R. Lüst, Z. f. Astrophysik, **51,** 163 (1961) and **57,** 192 (1963).

This indicates now that the material coming from the sun is *not* coupled to the sun in the sense that it would co-rotate. Complete co-rotation with the sun would mean that the material near the earth would have a velocity on the order of 450 km/sec. If the material emitted as corpuscular radiation from the sun would co-rotate at the earth's distance, we should observe a speed perpendicular to the radius of the order of 400 to 500 km/sec.

The fact that we are able to observe the second-order effect reflecting the individual motion of the comet, which is something like one-tenth or less than this, indicates that there is no co-rotation of the interplanetary plasma within, say, 20 km/sec or so.

This is mentioned specifically because it was proposed some time ago that the solar time variation of the intensity of cosmic radiation be ascribed to a co-rotation with the sun of the interplanetary plasma in the vicinity of the earth. It would seem that the evidence from the comets quite definitely rules out this hypothesis.

Of course, that leaves us with this question: Out to what distance do we have co-rotation with the sun? If we put together all the data on the electron density in interplanetary space as a function of the distance from the sun, we don't find any discontinuity. Of course, there are certain arguments discussed by R. Lüst, and by Lüst and Schlüter, which tend to show that, say, in the vicinity of the orbit of Mercury there is a surface beyond which there is no co-rotation and inside of which there may possibly be complete co-rotation as observed in the solar corona. This surface may be at a distance of, say, $\frac{1}{3}$ AU; but this estimate is highly uncertain. This is mentioned only as one of the questions that certainly should be answered by means of observations from space probes. One possibility would be to produce artificial plasma clouds much nearer to the sun and to make observations from a position that would permit us to see what is happening to them.

IONIZATION OF COMET TAILS

Another problem connected with these (Type I) plasma tails—the problem of the ionization—should be discussed briefly.

We mentioned the observable constituents, ions

like CO^+ and N_2^+ and some others: All of these have ionization potentials of the order of 14 or so electron volts.

Using Hinteregger's data to derive the time scale of photoionization, we get a result of the order of one to several months. That, of course, is by far too long to account for the observed rate of appearance of ion structures; therefore, if the best measurements known at present can be trusted, the ionization of these ions cannot be due mainly to the photoionization of the parent molecules (CO, N_2, $\cdots$).

That, then, leaves us with the problem of how we can account for the ionization, especially for the rapidity with which it sometimes occurs. Two or three possibilities will be mentioned.

The one discussed in the literature, in addition to photoionization, is the process known in atomic physics as *exchange of charge*. That is to say, a solar proton, in encountering a nonionized CO or N_2 molecule, may produce an ion of CO and becomes itself a neutral hydrogen atom; this happens with a large cross section if the velocity is in range of, say, 200 to 2000 km/sec. It has been found experimentally that the cross section depends strongly on the relative velocity. For the velocity range in question it is around 3×10^{-15} cm^2, a relatively large cross section.

The time scale of the order of 1 day (approximately 10^5 sec) for the formation of ionized structures would then require a particle flux of $10^{9.5}$ protons/cm^2-sec, provided the charge exchange takes place as in free interplanetary space.

Until a short time ago that seemed to be a satisfactory explanation. A particle flux of 3×10^9 protons/cm^2-sec would correspond, say, to 50 protons/cm^3 moving with a velocity of 600 km/sec. These figures were compatible with all the evidence available until about last year. The upper limits given by other means are around 100 particles/cm^3 for quiet conditions.

Now the measurements of Bridge and the similar experiments by the Russian physicists tend to indicate a flux of solar materials under quiet conditions of only several times 10^8 protons/cm^3-sec. This would correspond to a particle density on the order of 10/cm^3, and to a velocity on the order of, say, 300 km/sec and would leave for the flux a discrepancy by a factor of around 10.

Occasional observations indicate that the time scale of the formation of ion structures is as short as 1 hour; this means that we would have to have another $\frac{3}{2}$ powers of 10. But these last measurements are probably characteristic for disturbed conditions, that is, to higher solar activity.

To account for the ionization, there are two possibilities in addition to charge exchange mentioned in current discussions or in the literature: One is *ionization by electron collisions*, while the other one is a *binary chemical reaction* of the following type: starting, say, with $C^+ + O_2 \rightarrow CO^+ + O$, which is both exothermic and has a relatively large cross section.

With regard to electron collisions, there is the difficulty that we need electrons with energy of at least something like 25 eV because the cross section for ionization is quite low for electrons with smaller energy.

With such electron energies we would expect, first, collisional dissociation rather than ionization, with quite a number of competing processes in part of which negative ions are formed. So it would seem that the chances that electron collisions contribute efficiently to ionization are not very good.

The other possibility has some slight advantages: The ionization potential of C is only 11.63 eV; hence the photoionization should be somewhat faster than that of CO.

The ionization potentials of CN or C_2 are not known, but we might guess that they would be similar to that of C. Unfortunately the photoionization cross sections for these are, as far as we know, not yet available. Summing up, it may be said that the cause of the ionization could be complex although, on the balance, the mechanism of charge transfer is likely to be the most efficient contributor.

ARTIFICIAL PLASMA CLOUDS

Lastly, a few words should be said about the possibility and usefulness of producing artificial plasma clouds in interplanetary space to obtain more information about the properties of solar corpuscular radiation by direct observation.

Evidently what we would like to have is a material that, in ionized form, has at least one fairly strong transition in the ordinary optical

range and at the same time becomes photoionized by ordinary solar light as rapidly as possible.

If we look at the periodic table we find easily that, practically, only some of the alkali earth atoms are really useful.

Next we inquire what quantities are needed in order to enable us to observe each clouds reasonably well without extra techniques, such as special filters or image converters—which of course would considerably increase the possibilities of this type of work.

With this narrow restriction, we have worked out the question in some detail. However, we will not go into this, since a paper containing all this information will appear quite soon in the *Zeitschrift für Astrophysik* (see vol. 53, p. 226).

The minimum final figures needed for an assumed cloud distance of 200,000 km are: for calcium, 2 kg; for strontium, 6 kg; while, for barium, only 130 gm.

The reason that we require a rather large quantity of calcuim is that the main calcium resonance absorption line is quite strong in the solar spectrum itself.

Also, the strontium line is at a disadvantage by its position in the solar spectrum; only the barium resonance doublet (around 4600 and 4900 Å) is really in a very good region of the solar spectrum. This and the fact that the photoacceleration of the barium ion is not large (6 cm/sec^2) overbalance the disadvantage that barium has the larges atomic weight (135).

So the general technique would be to release a quantity of barium gas (approximately 100 gm or more at a distance of 200,000 km) and to see how the solar plasma interacts with it. This would give information directly about the direction of acceleration; it would also give indirect information about the mechanism of coupling if we use, say, a mixture of barium and calcium, where the large difference in atomic weights might discriminate between mechanisms. Such a difference would help us to distinguish between, for instance, plasma acceleration and the influence of magnetic fields.

The sideways spreading of the material should give information on the interplanetary magnetic fields, but these details will not be discussed here.

In deriving the figures given so far, it was assumed that the surface brightness of the cloud should be greater than the integral surface brightness of the sky—a condition that would have to be fulfilled when ordinary observing techniques are used. But, by applying a filter, we should gain some advantage.

With techniques of this sort, we could furthermore hope to use not only alkali earths but also carbon monoxide—but only during periods of very much enhanced solar particle flux. If we would make such an experiment during a big magnetic storm, we should indeed be able to observe directly the ionization of the CO by exchange of charge. Since we know the cross section and we know precisely the amount that we have released, we should get the time scale of the ionization just by watching the appearance of this carbon monoxide ion. Knowing the quantity involved, we should be able to make an independent quantitative check of the intensity of the solar particle stream.

This, of course, is only of interest if the experimental measurements with plasma probes appear to be so difficult that we would like to have a check done by an entirely independent method. If, on the other hand, we are confident that the available experimental techniques are perfectly adequate, then perhaps there is no particular point in making an experiment with CO since it might not be regarded as worthwhile from our interest in the physics of comets.

CONCLUDING REMARKS

Let us conclude with one general remark. To the extent we succeed in understanding what is going on in the comets, specifically in the plasma tails of comets, we can use these tails as probes provided by nature. That has, of course, an advantage, because then we can use all the past observations of comets which have appeared in the literature. We mentioned the comets moving above the poles of the sun and the fact that they appear in all phases of the solar cycle. So, in spite of the fact that we are fairly confident in the near future of a number of space probes and very distant satellites to give us direct measurements, it might still be worthwhile to have additional clues from this very rich material in the form of

past observations of comets. These data certainly would give us information of a somewhat different character—for instance, if we inquire about statistical properties of solar corpuscular radiation over long periods of time, or its properties at high heliographic latitudes, or in a part of the solar cycle for which we shall now have to wait 10 years. In this connection the intended experiment of Dr. Donn of Goddard Space Flight Center is very much welcomed, although its aim is somewhat different from that of our own experiment. It should help us also to understand more in detail the processes which take place on comets, and thus, though perhaps somewhat more indirectly, assist us in using the comets as probes for the interplanetary plasma.

Appendix A

DISCUSSION

Q. As you undoubtedly know, the Russians injected a couple of plasma clouds at fairly large distances from the earth. These were observed for periods of several minutes which, judging from the acceleration you gave, might—the motion, say, due to solar plasma might—be observable in the periods when solar clouds were observable. Were useful observations of those clouds made?

Dr. Biermann: I am not aware of observations of plasma clouds. I have heard only of observations of artificial natrium clouds.

Q. Sodium clouds?

Dr. Biermann: Yes. But the observed clouds are not ionized. When ionized, sodium becomes invisible. So what is observed is the nonionized component, and that is analogous to the CN and C_2, but not to the plasma. The nonionized component doesn't give you any information about the solar plasma. It is simply acted on by solar light. Whatever influence the plasma can exert on the nonionized component would be quite small compared with the light pressure.

Q. If the ionization of CO and N_2 occur through charge exchange, is there any reason why C_2 and CN cannot be ionized in the say way? CN, I think, would be visible—much better than C_2. C_2, I think, wouldn't be visible in the low-lying state.

Dr. Biermann: In this particular case I was rather thinking of photoionization, and not of ionization by exchange of charge. Of course, it might very well be that charge exchange also enters significantly. That I don't know. But the cross section for exchange of charge depends in a fairly complicated way on the properties of the spectrum. It seems that the fact that both CO^+ and N_2^+ have low-lying levels giving rise to the observable resonance terms may have some connection with the fact that the cross section for exchange of charge is large—in the velocity range of a few hundred to a few thousand kilometers per second [see the work by Gilbody and Hasted where further references may be found; *Proc. Roy. Soc.* A238: 334 (1956)].

I am not aware of any resonance lines of nonionized CO in the optical range. I was also under the impression that the ion of CN does not have any observable resonance bands. If I am wrong there, then of course it would be very significant in connection with the questions you are discussing.

Q. You mentioned the fact that there is a rotation observed in tails sometimes. Is this possibly explainable in terms of a longitudinal drift of particles in a field gradient?

Dr. Biermann: The way we tend to look at it is this: You know perhaps of the work of Lüst and Schlüter on force-free fields. First, Lundquist, then Lüst and Schlüter, and later still others discussed the properties of force-free fields. One of the main properties is that they have a sort of spiral structure if there is any symmetry at all.

Now, we do observe occasionally a feature that I have not mentioned, something like a large-scale helical structure in the tail (type I) of a comet. The tail then looks like a large helix viewed from the side. Of course, that is not an unambiguous interpretation, but it is one that we would naturally think of. We have liked this idea, even since we became aware of the probability that there were magnetic fields that should be force-free.

Of course, the situation might be more complicated than that.

Q. As an extension of this, is there any in-

formation about whether there is a hole in the magnetic field behind the comet? Whether the solar field is excluded by the comet?

Dr. Biermann: There is no information concerning the magnetic field; but there is an indication, which I had not mentioned, as the particle stream. The indication is this: If the acceleration is measured as a function of position in the tail, it is often found (e.g., in the work of Peter Stumpff on comet Morehouse and in that of Rhea Lüst on comet Mrkos) that the acceleration strongly decreases, possibly to about zero in the vicinity of the tail axis at some distance behind the coma.

So it looked as though in these cases the influence of the interaction of the solar corpuscular radiation with the comet's coma was of such a kind that the immediate vicinity of the axis of the tail was excluded from further impact of solar corpuscular radiation.

NUCLEAR COMPOSITION AND RIGIDITY SPECTRA OF SOLAR COSMIC RAYS

S. BISWAS* AND C. E. FICHTEL

Goddard Space Flight Center

Further studies on the composition of solar cosmic rays have been made particularly with regard to the nuclei of charge 10 or more. The results support the earlier conclusion that the composition of the multiply charged nuclei in solar cosmic rays seems to reflect that of the sun as determined by spectroscopic means for those nuclei where comparisons can be made. The relative abundances of helium and neon which cannot be determined well in the photosphere by spectroscopic means are found to be 107 ± 14 and 0.13 ± 0.02 relative to oxygen, respectively, for the energetic solar particles. The helium nuclei rigidity spectrum during one period in the November 12, 1960, solar cosmic ray event has been extended to lower rigidities than measured previously, thereby permitting a comparison to the proton rigidity spectrum over a wider interval. The result of this analysis together with previous work demonstrates that it is unlikely that any of the simple spectral forms proposed in the literature yield a consistent answer for both protons and helium nuclei for all events.

INTRODUCTION

The energetic nuclei of solar origin arriving at the earth after some solar flares are known to consist primarily of hydrogen nuclei. However, recent investigations have revealed the presence of particles other than protons. Solar helium nuclei have been measured in a number of events in nuclear emusions flown on balloons (Biswas, Freier, and Stein, 1961; Ney and Stein, 1962; Freier and Webber, 1963) and in sounding rockets (Biswas, Fichtel, and Guss, 1962; Biswas, Fichtel, Guss, and Waddington, 1963). Heavy nuclei of solar origin have been measured with nuclear emulsion stacks in six sounding-rocket firings during three solar cosmic ray events (Fichtel and Guss, 1961; Biswas, Fichtel, and Guss, 1962; and Biswas, Fichtel, Guss, and Waddington, 1963) (the two papers published in 1962 and 1963 by Biswas et al. are hereinafter referred to as "paper I" and "paper II," respectively). Detectors flown in satellites (Pomerantz and Witten, 1962; Yagoda, Fitz, and Fukui, 1961) have also indicated the presence of solar-heavy nuclei. In the work presented in papers I and II, the relative abundances of the hydrogen, helium, carbon, nitrogen, oxygen, and heavier nuclei were obtained for three different solar events, and an upper limit was set for the abundance of the light nuclei ($3 \leq Z \leq 5$) in each case. The composition of the multiply charged component of the solar cosmic rays was found to be the same for each event within uncertainties. Further, it seemed to be a reflection of the composition of the solar surface, as determined by spectroscopic means.

In this paper we shall present further work on the heaviest nuclei for which measurements could be made, describe additional data obtained on the helium nuclei energy spectrum from the November 12, 1960, event which extends the helium nuclei spectrum to lower energies than reported previously, and then summarize the existing data on the nuclear composition and the energy spectra of the various components. The significance of these results in terms of galactic cosmic rays and solar composition will also be considered briefly.

*NASA—National Academy of Sciences senior postdoctoral resident research associate on leave of absence from the Tata Institute of Fundamental Research, Bombay, India.

NEW EXPERIMENTAL RESULTS

The data upon which the analysis is based were obtained from measurements on particle tracks in nuclear emulsions flown aboard sounding rockets fired from Fort Churchill, Canada. The two particular flights from which the results reported here were obtained were made at 1840 U.T. on November 12 and 1630 U.T. on November 13, 1960. There was a low-energy cutoff for the particles included in the analysis determined by the material above the emulsion, 0.19 gm/cm^2, and the minimum length in the emulsion accepted for analysis. Because of the very steep energy spectrum, most of the solar nuclei observed stopped within a centimeter of the minimum accepted range. For definiteness, however, a maximum energy of about 130 MeV/nucleon was set. The following discussion of the method of data reduction used in this experiment has been kept brief because the experimental procedures were very similar to those described earlier in papers I and II.

Heavy Nuclei

Because of the low rate of occurrence of the elements with nuclear charges greater than that of oxygen, there were only a few of these particles in the samples obtained, except for the flights in the November 12, 1960, event wherein the tracks of an appreciably larger number of nuclei were recorded. The original report on the composition of the November 12, 1960, solar event was based on a total of 311 heavy nuclei ($Z \geq 3$), but most of these were in the carbon, nitrogen, oxygen group and only 24 had higher charges. The reduction of the November 12, 1960, data has now been extended in order to obtain a more exact determination of the relative abundance of these heavier nuclei.

In this extended work, a total of 125 new tracks was found which could be attributed to nuclei heavier than oxygen in the flight at 1840 U.T. on November 12. Of these about half were formed by neon nuclei and the remainder were of higher charge. The relative abundances were determined

TABLE 1.—*Relative Abundances of Nuclei Normalized to a Base of* 1.0 *for Oxygen*

Element	Solar cosmic rays*	Sun†	Universal abundances§	Galactic cosmic rays‡
$_2$He	107±14	(?)	150	48
$_3$Li	----------	≪0.001	≪0.001	0.3
$_4$B–$_5$B	<0.2	≪0.001	≪0.001	0.8
$_6$C	0.59 ±0.07	0.6	0.3	1.8
$_7$N	0.19 ±0.04	0.1	0.2	≤0.8
$_8$O	1.0	1.0	1.0	1.0
$_9$F	<0.03	≪0.001	≪0.001	≤0.1
$_{10}$Ne	0.13 ±0.02	(?)	0.40	0.30
$_{11}$Na	----------	0.002	0.001	0.19
$_{12}$Mg	0.043±0.011	0.027	0.042	0.32
$_{13}$Al	----------	0.002	0.002	0.06
$_{14}$Si	0.033±0.011	0.035	0.046	0.12
$_{15}$P–$_{21}$Sc	0.057±0.017	0.032‖	0.027	0.13
$_{22}$Ti–$_{28}$Ni	≤0.02	0.006	0.030	0.28

*See text for a discussion of the manner in which the values in this column were determined.

†The uncertainty of the values in this column is probably of the order of a factor of 1.5 to 1 (see Aller, 1961, or Goldberg, Muller, and Aller, 1960).

‡The uncertainty of the values in this column is hard to estimate, but is probably at least a factor of 2 in some cases (see Suess and Urey, 1956, and Cameron, 1959).

§The uncertainty of the values in this column varies from 10 to about 30% (see Waddington, 1960).

‖A 5/2 ratio for the abundance of $_{16}$S relative to $_{18}$A was assumed, the relative abundance of $_{18}$A being unknown.

TABLE 2.—*Helium Nuclei Fluxes for Flight* 10.16 *at* 1603 *U.T. Nov.* 13, 1960

Energy interval MeV/nucleon	dJ/dE part/cm³ sr sec MeV)	dJ/dR part/cm² sr sec MeV)
15.6–18.5	17.1±5.4	1.60±0.50
18.5–20.5	18.6±5.6	1.86±0.56
20.5–23.5	9.0±3.2	0.93±0.33
23.5–29.0	4.3±1.6	0.48±0.18

from ratios of fluxes in the same energy-per-nucleon intervals. For nuclei with a charge, Z, less than 8, the energy-per nucleon interval was 42.5 to 135 MeV/nucleon, and for nuclei with $Z>8$ the lower limit gradually increases, since the range of a nucleus for a given energy per nucleon is a decreasing function of the quantity Z^2/M. For example, the relative abundance of $_{12}$Mg was determined by a comparison of the fluxes of $_{12}$Mg and $_{8}$O in the interval from 53.5 to 135 MeV/nucleon.

Due to limitations of the experimental technique and the very short tracks involved, it is not possible to resolve the charges of the particles unambiguously above about charge 10. However, there is a large even-charge excess over odd charges existing in the sun in the region of the charge spectra from neon to calcium. If this preponderance is assumed in the solar cosmic rays also, the nuclei can reliably be assigned the even charge which they fall nearest on the basis of the measurements. The results are summarized in table 1 along with previous results.

Helium Nuclei

In paper I the flux and energy spectrum of solar helium nuclei were obtained in the energy interval from 29 to 130 MeV/nucleon during the second rocket flight at 1603 U.T., November 13, 1960. In this work, the flux and energy spectrum of solar helium nuclei were measured in the energy interval from 15 to 29 MeV/nucleon during the same rocket flight. The details of the experimental analysis are given in the appendix, and the results are given in table 2. The intensities and rigidity spectra of solar protons and helium nuclei can now be compared over a wider interval. A discussion of the results and significance of this comparison will be given in the next section.

GENERAL EXPERIMENTAL RESULTS

We shall begin this section by summarizing the present and the previous data on the relative abundances and energy spectra of solar hydrogen, helium, and heavy nuclei. Beginning with the composition of the multiply charged component, it has been shown in papers I and II that the helium and medium nuclei had the same energy per nucleon spectrum, and hence the same velocity and rigidity spectra. For this reason and theoretical considerations related to the acceleration and propagation of particles discussed in papers I and II, it seems reasonable to present the relative abundances in terms of the same kinetic energy-per-nucleon interval. When the abundances are determined in this way, the composition is found to be the same within uncertainties for all the measurements that have been made (see paper II). Therefore, there is good reason to take the average values of all the measurements as the best estimate of the relative abundances of these elements in solar particle events.

Table 1, which combines the previous data with those of this paper, shows that for those elements where a comparison can be made to spectroscopic measurements in the photosphere, the composition of the multiply charged nuclei seems to be at least similar and may be identical. For comparison, the relative abundances of these same nuclei in the universe and in galactic cosmic rays are shown. It is seen that the galactic cosmic ray composition is very different from the rest. The data in table 1 confirm the earlier results and show the significant differences between the composition of the solar and galactic cosmic rays. The differences are most striking with respect to the relative abundances of the heavier elements, for example, the helium to medium nuclei ($6\leq Z\leq 9$) ratio is 60±7:1 in the solar cosmic rays as compared with 14:1 in the galactic cosmic rays, and the ratio of helium nuclei to nuclei with charges $11\leq Z\leq 28$ is 803±93:1 in the solar cosmic rays as compared to approximately 48:1 in the galactic cosmic rays. The other outstanding difference, the

light ($3 \leq Z \leq 5$) to medium nuclei ratio of >0.001 in solar cosmic rays as compared to 0.3 in galactic cosmic rays may be ascribed to the fact that the high-energy galactic cosmic rays have traversed about 3 gm/cm^2 of interstellar matter whereby light nuclei are formed by the fragmentation of heavier nuclei. The unusual composition of galactic cosmic rays is not entirely understood, but may be explained in terms of supernovae origin (Ginsburg and Syrovatsky, 1961).

Turning now to the energy distribution, of the many spectral representations we have chosen the following two for purposes of comparing the protons and multiply charged nuclei:

$$dJ/dW = C(W^{-\gamma}/W_0) \qquad (1)$$

and

$$dJ/dR = A \exp(-R/R_0), \qquad (2)$$

where J is the integral particle flux, W the total energy per nucleon, W_0, the rest energy per nucleon, and R the particle rigidity ($R = pc/Z$, where p = momentum). It has been shown in paper II that form (1) reduces to form (2) for kinetic energies >35 MeV and γ-values commonly occurring in solar proton events. Equation (2) seems to be a good representation down to energies as low as approximately 10 MeV, at least for the middle and late phases of solar events (Freier and Webber, 1963).

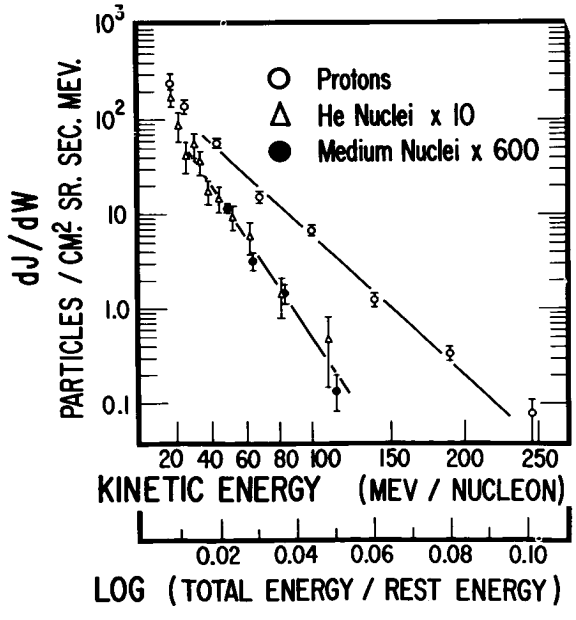


FIGURE 1.—The differential energy per nucleon spectra of hydrogen, helium, and medium nuclei at 1603 U.T., November 13, 1960. ○ = protons; △ = He nuclei × 10; ● = medium nuclei × 600.

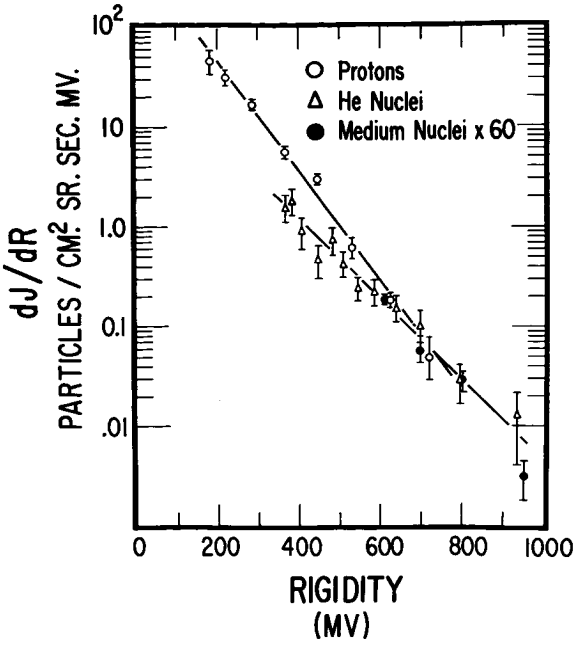


FIGURE 2.—The differential rigidity spectra of hydrogen, helium, and medium nuclei at 1603 U.T., November 13, 1960. ○ = protons; △ = He nuclei; ● = medium nuclei × 60.

In figure 1 we have shown the differential energy spectrum of solar helium nuclei in the energy interval from 15 to 130 MeV/nucleon in the rocket flight at 1603 U.T. on November 13, 1960, by combining the present results with those obtained previously. The differential energy spectra of protons and medium nuclei measured in the same flight are also shown. The rigidity spectra of solar hydrogen, helium, and medium nuclei are shown in figure 2. The values of γ and R_0 have been calculated by the least-squares method under the assumption that the mathematical expression was a satisfactory representation of the experimental spectrum within the uncertainties. These values, along with the associated error estimates are given in table 3.

Similarly the values of γ and R_0 were calculated for protons, helium, and medium nuclei obtained previously in the rocket flights on November 12,

TABLE 3.—*Values of* γ *and* R_0

	1840 U.T.* Nov. 12, 1960	1603 U.T.* Nov. 13, 1960	1951 U.T.* Nov. 16, 1960	0600 U.T.* Nov. 17, 1960	0339 U.T.* Nov. 18, 1960	1408 U.T.† Sept. 3, 1960
Experimental values for γ of equation (2)						
Protons	−20.7±2.4	−37.4±2.2	−26.5±3.4	−32.4±2.1	−36.7±2.6	−11.9±0.4
Helium nuclei	−56.3±7.2	−59.7±8.2	−70.8±10.8	−63.8±9.9	−60.8±8.4	−27.0±2.1
Medium nuclei ($6 \leq Z \leq 9$)	−61.8±9.6	−64.4±8.9	−63.8±8.4	−71.0±11.0	−60.7±12.3	----------
Experimental values for R_0 of equation (3), in MeV						
Protons	+119.3±7.0	+78.0±3.9	+103.3±10.0	+89.7±5.7	+81.3±6.4	+169.6±6.9
Helium and medium nuclei (average)	+103.8±9.9	+109.2±6.7	+107.2±10.4	+102.9±12.4	+113.9±15.5	+132.6±7.2‡

*Data from papers I and II, and this article.
†Data from Fichtel and Guss (1961); Biswas, Freier, and Stein (1961), and paper II.
‡Helium nuclei data only.

15, 16, and 18. These are also shown in table 3. The energy spectrum of the protons is in general measured over a wider interval. During the September 3, 1960, rocket flight, energy spectra for solar protons and helium nuclei were measured at balloon altitude at Minneapolis by Biswas, Freier, and Stein (1961). The Minneapolis balloon data for particles of rigidity >0.95 BeV/c can be compared to those obtained in the rocket flight from Fort Churchill, as discussed by Biswas, Freier, and Stein (1961) and Winckler, Bhavsar, Maseley, and May (1961). Therefore, the balloon data for particles of rigidity >0.95 BeV/c have been combined with the rocket data (Davis, Fichtel, Guss, and Ogilvie 1961) to obtain the energy spectra of protons from 15 to 1000 MeV and of He nuclei from 45 to 210 MeV/nucleon as shown in figure 3. The corresponding rigidity spectra are shown in Figure 4.

From the results shown in figures 1–4 and table 3, the following conclusions can be reached:

1. In the five instances where the energy spectra of the helium and medium nuclei could be measured the value of γ was the same for He and medium nuclei within experimental errors. On the other hand proton and He nuclei have very different values of γ in general. This result is a strong argument in favor of the hypothesis that helium and medium nuclei are ionized in such a way that they have the same charge to mass ratio during the acceleration and propagation phases. Because of the very different ionization potentials the only reasonable conclusion is then that they are completely ionized.

2. Comparing the rigidity spectra and the R_0 values of protons and multiply charged particles, it is seen that, although they are in general similar, they are not the same. The difference in R_0 values is particularly significant in the two flights where $Z \geq 2$ data are available over a wider rigidity interval. Out of the six events for which measurements could be made, R_0 values differ by less than 1 standard deviation (σ) in one case, 1σ to 2σ in two cases, 2σ to 3σ in one case, and more than 3σ in two cases. The probability of deviations of this order is less than 0.1 percent, even if the September 3, 1960, result is not included. The rigidity spectra of protons and

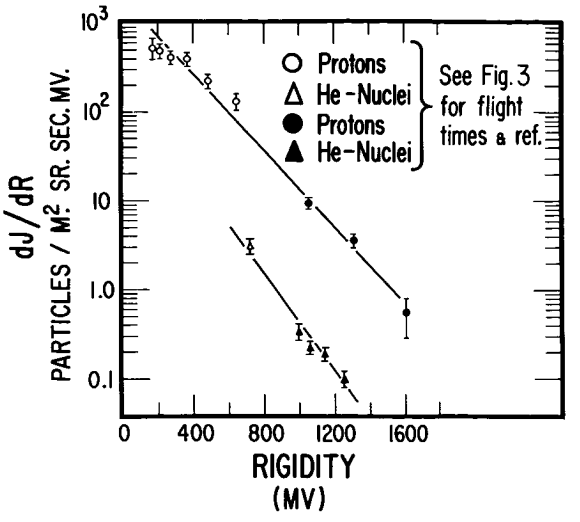


FIGURE 4.—The differential rigidity spectra of hydrogen and helium nuclei at about 1400 U.T., September 3, 1960. ○ = protons, 1408 U.T., September 3, 1960 (Davis et al., 1961); △ = He nuclei × 10, 1408 U.T., September 3, 1960 (Biswas et al., 1963); ● = protons, 0930–1700 U.T., September 3, 1960 (Biswas et al., 1961); ▲ = He nuclei × 10, 0930–1700 U.T., September 3, 1960 (Biswas et al., 1961).

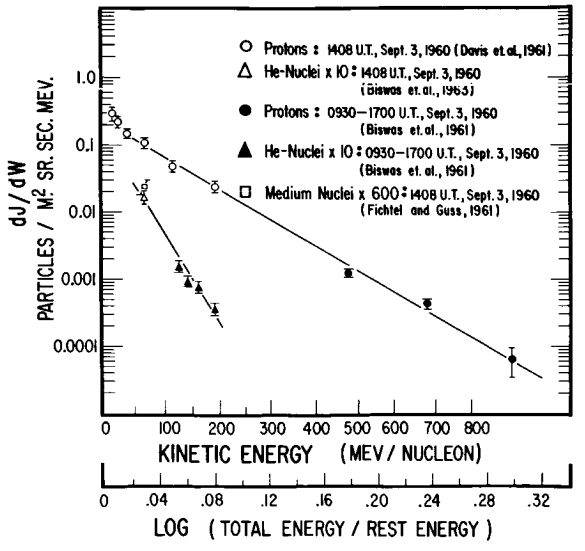


FIGURE 3.—The differential energy per nucleon spectra of hydrogen and helium nuclei at about 1400 U.T., September 3, 1960. ○ = protons, 1408 U.T., September 3, 1960 (Davis et al., 1961); △ = He nuclei × 10, 1408 U.T., September 3, 1960 (Biswas et al., 1963); ● = protons, 0930–1700 U.T., September 3, 1960 (Biswas et al, 1961); ▲ = He nuclei × 10, 0930–1700 U.T., September 3, 1960 (Biswas et al., 1961); □ = medium nuclei × 600, 1408 U.T., September 3, 1960 (Fichtel and Guss, 1961).

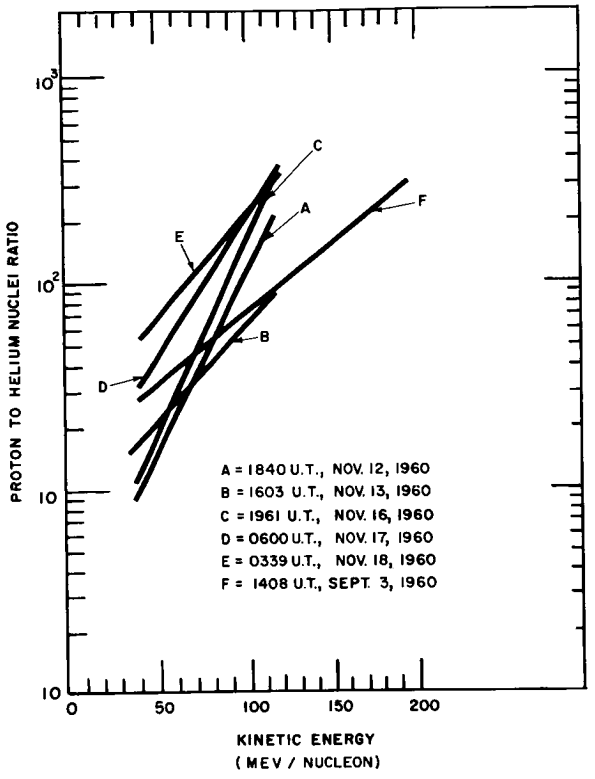


FIGURE 5.—The proton-to-helium nuclei ratio as a function of kinetic energy per nucleon at several different times. In each case, the line represents a best-fitting straight line to the existing data. Uncertainties in the ratios range from 25 to 50%. A = 1840 U.T., November 12, 1960; B = 1603 U.T., November 13, 1960; C = 1961 U.T., November 16, 1960; D = 0600 U.T., November 17, 1960; E = 0339 U.T., November 18, 1960; F = 1408 U.T., September 3, 1960.

multiply charged nuclei are therefore probably not the same. Another feature is that even though the R_0 values for protons and He nuclei are similar, the ratio of protons to He nuclei, or protons to medium nuclei, in the same rigidity interval varies greatly from one time to another in a given event and from one event to the next. Summaries are given in table 2 of Freier and Webber (1963) and table 8 of paper II.

3. Although the values of γ are very different for singly and multiply charged particles, the proton-to-helium, or proton-to-medium, ratio in a given kinetic energy per nucleon interval is more nearly the same. Again, however, the variations are sufficiently large so that the exact agreement is in fact very improbable (see table 8 in paper II). The proton-to-helium ratios for several events are plotted as a function of energy per nucleon in figure 5. Notice the tendency for the proton-to-helium ratio to vary less with kinetic energy per nucleon late in an event.

DISCUSSION

Beginning with the composition of the multiply charged particles, we shall first summarize the reasons for believing that among the multiply charged nuclei where the charge-to-mass ratio is essentially the same for all nuclei, the composition measured in solar particles may be the same as that in the sun. First, they have had the same relative abundances within uncertainties each time a measurement was made, whereas this has not been true for the protons and helium nuclei which have different charge-to-mass ratios. Second, their composition is the same as that of the sun insofar as comparisons can be made. Third, the helium and medium nuclei have had the same energy per nucleon spectrum in the two events where measurements could be made at the same time, whereas the proton and helium nuclei spectra in the same events were very different. The observations mentioned above make a biased injection unlikely, and, after that step in addition to the experimental evidence, it is theoretically unlikely for a bias to occur in the main acceleration or propagation phase, since the particles have the same charge-to-mass ratio.

If it is assumed that an unbiased sample of the multiply charged nuclei discussed has been obtained, there is now an estimate of the relative abundances of solar neon and helium, both of which cannot be determined in the photosphere by normal spectroscopic techniques. The neon-to-oxygen ratio is similar to the universal abundances estimated by Suess and Urey (1956) and Cameron (1959), although a bit on the low side. The helium-to-medium ratio is typical, but, of course, the ratio of hydrogen-to-helium nuclei is of more interest. For the reasons mentioned above, associated with the different energy spectra and charge-to-mass ratio, there is no simple reliable way to determine this ratio from solar cosmic rays alone. If the helium-to-medium ratio of 60 ± 7 is accepted as representative of the sun and the proton-to-medium value from spectroscopic data,

namely, 650 (Aller 1961; Goldberg, Muller, and Aller 1960), is used, a proton-to-helium ratio of 11 is obtained. The uncertainty in this number depends on the correctness of the assumption above and the uncertainty in the proton-to-medium ratio. Unfortunately, therefore, this estimate of the proton-to-helium ratio is probably good only to within a factor of 2.

The results of the second half of this work, especially when combined with earlier work, indicate strongly that, although the exponential rigidity spectrum, equation (2), seems to fit the experimental data for solar protons and helium nuclei quite well, at least above about 10 MeV, the values of R_0 for protons and helium nuclei are not the same, at least not always. This result together with the earlier discussions of spectral shape in other works (e.g., Freier and Webber 1963) indicates that it is unlikely that any of the simple spectral forms proposed in the literature are consistent with the experimental data for both protons and helium nuclei. However, since the propagation of solar particles almost certainly depends on both the particle's velocity and rigidity, this conclusion is not inconsistent with our present limited understanding of the acceleration and interplanetary propagation of energetic solar particles.

Appendix

HELIUM NUCLEI DATA REDUCTION

In order to measure the flux and energy spectra of solar helium nuclei at lower energies than reported previously (paper I), the scan line for the flight on November 13 was set closer to the outer edge of the emulsion, specifically at 0.60 mm from the periphery. All particles crossing the scan line in the middle half of the emulsion thickness were recorded if they were within the solid angle defined by the azimuthal angle $\alpha \leq 45°$ and dip angle $\beta \leq 7.1°$, and had an ionization ≥ 28.4 times minimum (grain density ≥ 62 grains/100 μ). With these criteria, the lower limit to the ambient energy of the helium nuclei which could reach the scan line after traversing the shielding and emulsion (equivalent to 1.20 mm of emulsion) was 15.6 MeV/nucleon. The upper limit to the ambient energy of the helium nuclei recorded was 36.5 MeV/nucleon. However, due to the statistical fluctuation of the ionization measurements, the actual upper limit below which an unbiased sample was obtained was 29 MeV/nucleon.

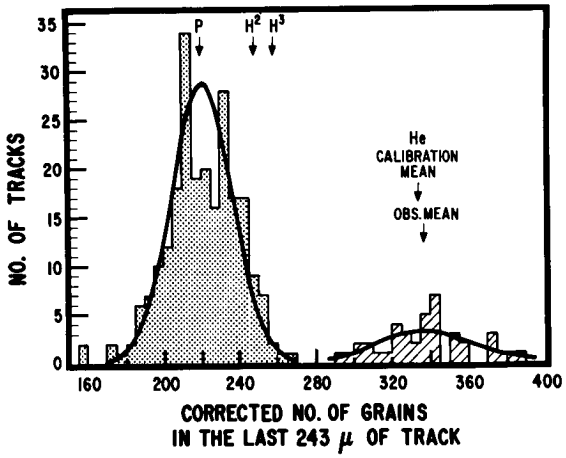


FIGURE 6.—The distribution of the number of grains in the last 243 μ from the end of 274 particle tracks showing the resolution of proton and helium nuclei. The smooth curves show the Gaussian distribution normalized to the number of tracks with standard deviation taken as that expected from the combined statistical and systematic error, 7.4%. The expected positions of the mean values for deuteron, triton, and helium nuclei tracks are also shown.

To resolve helium nuclei from protons, all accepted particle tracks were followed to the end of their range and the integral number of grains in the last 243 μ from the end was measured. It was found that the integral number of grains in this distance was adequate to resolve these particles because the Ilford G5 emulsion had been underdeveloped to such an extent that the minimum ionization corresponded to 2.6 grains/100 μ. Since the grain counting was made close to the edge of the emulsion however, adequate care was taken to normalize the grain count for the variation with distance from the edge and with depth in the emulsion. These variations were determined by using identified helium tracks obtained in the previous work, and were similar to those shown in figures 3 and 4 of paper I. These corrections were, in general, small except for some extreme cases. Since all measurements were confined to a single emulsion, the problem of plate-to-plate normalization did not arise.

The distribution of the corrected number of grains in the last 243 μ of range is shown in figure 6 for the 274 particle tracks used in the analysis. It is seen that the 40 helium nuclei tracks are completely resolved from the proton ones. The energy per nucleon of helium particles was determined from total range measurements. The flux in a given energy interval was calculated from the known trajectory of the rocket and its orientation with respect to the zenith in the same manner as described previously in paper I.

REFERENCES

ALLER, L. H., *The Abundance of the Elements* (New York: Interscience Publishers), 1961.

BISWAS, S., FICHTEL, C. E., and GUSS, D. E., *Phys. Rev.*, **128,** 2756 (Paper I), 1962.

———, ———, ———, and WADDINGTON, C. J., *J. Geophys. Res.*, **68,** 3109 (Paper II), 1963.

———, FREIER, P. S, and STEIN, W., *J. Geophys. Res.* **67,** 13, 1961.

CAMERON, A. G. W., *Ap. J.*, **129,** 676, 1959.

DAVIS, L. R., FICHTEL, C. E., GUSS, D. E., and OGILVIE, K. W., *Phys. Rev. Letters*, **6,** 492, 1961.

FICHTEL, C. E., and GUSS, D. E., *Phys. Rev. Letters*, **6,** 495, 1961.

FREIER, P. S., and WEBBER, W. R., *J. Geophys. Res.*, **68,** 1605, 1963.

GINZBURG, V. L., and SYROVATSKY, S. I., *Prog. Theoret. Physics Suppl.*, No. 20, p. 1, 1961.

GOLDBERG, L., MULLER, E. A., and ALLER, L. H., *Ap. J. Suppl.*, **5,** 1–138, 1960.

NEY, E. P., and STEIN, W., *J. Geophys. Res.*, **67,** 2087, 1962.

POMERANTZ, M. A., and WITTEN, L., *3d Int. Space Sci. Symp. of COSPAR*, Washington, D.C., May 1–9, 1962.

SUESS, H. E., and UREY, H. C., *Rev. Mod. Phys.*, **28,** 53, 1956.

WADDINGTON, C. J., *Prog. Nuclear Phys.*, **8,** 3, 1960.

WINCKLER, J. R., BHAVSAR, P. D., MASSELEY, A. J., and MAY, T. C., *Phys. Rev. Letters*, **6,** 488, 1961.

YAGODA, H., FITZ, R., and FUKUI, K., *Phys. Rev. Letters*, **6,** 626, 1961.

COSMIC RAYS OBSERVATIONS IN SPACE

D. A. BRYANT,* T. L. CLINE, U. D. DESAI* AND F. B. McDONALD

Goddard Space Flight Center

The Explorer XII cosmic ray experiment is described and measurements made during the solar event of September 28, 1961, are discussed. Galactic cosmic ray measurements are also reported. A few hours before the class 3 flare of September 28 two short counting rate increases were observed and these have been interpreted as electron bursts. The anisotropy of the solar protons is described. It is found that the history of the intensity of the solar protons is consistent with their having diffused through interplanetary space with an effective mean free path of 0.04 AU. An estimate of the distance from the sun at which diffusion becomes unimportant and particles escape gives 2-3 AU. Simple diffusion, does not account for the anisotropy observed early in the event. Two days after the flare there was a large increase in the intensity of protons with energies less than 15 Mev. As most of these particles, which have been called "energetic storm particles," arrived after the sudden commencement, it is suggested that they were solar protons trapped within the plasma cloud which caused the magnetic storm. A possible trapping mechanism is outlined. Explorer XII measurements of the Forbush decrease of September 30, 1961, are compared with neutron monitor measurements at Deep River. The decrease is larger at Explorer XII by a factor of 1.7 ± 0.3.

INTRODUCTION

Explorer XII (1961υ) was launched on August 16, 1961, into a highly elliptical orbit having an initial perigee of 6700 km, an apogee of 83,600 km, and an orbital period of 26.5 hours. During its active life the satellite, at apogee, was on the sunlit side of the earth. For more than half the time in each orbit the satellite was beyond the magnetosphere so that it was possible to obtain cosmic ray measurements free from the influence of the earth's magnetic field and the Van Allen radiation.

The cosmic ray experiment on Explorer XII was designed to measure the intensity and energy spectra of the galactic cosmic ray protons between 100 and 600 Mev and the total proton intensity above 600 Mev. The objectives of the experiment were to study the 11-year modulation, diurnal variations, Forbush decreases, and other modulations that occur in the primary cosmic ray intensity in addition to the dynamics of solar cosmic ray events.

*NAS–NASA Resident Research Associate.

The lower limit of the detectable proton energy spectrum can be extended from 100 Mev, the quiet-time value, to 3 Mev during solar cosmic ray events. This extension, possible only when the intensity at low energies exceeds the background produced by the high-energy particles, provides a sensitivity about 50 times greater than that which can be achieved by riometers.

This paper describes the detectors used and discusses the measurements made during the solar cosmic ray event initiated by a class 3+ flare on September 28, 1961. Fortunately, the satellite was at apogee both at the time of the arrival of the solar cosmic rays and when the magnetic storm began two days later.

The rise and recovery of the solar proton intensity were recorded as a function of energy from 3 to 600 Mev. A plasma cloud, apparently emitted at the time of the flare, produced a large magnetic storm and a moderate Forbush decrease about 46 hours later. At the time of the sudden commencement of the magnetic storm, an increase in the intensity of low-energy ($E<30$ Mev) protons was observed. The other four solar cosmic

ray events observed during the 112-day active life of the satellite, and further details of the modulation of galactic cosmic rays, will be discussed in subsequent reports.

DETECTORS

Three cosmic ray detectors were used. The details of the electronics for this experiment have been described by Desai, Porreca, and Van Allen.[1] A scintillation counter telescope (fig. 1) was used to detect medium-energy ($E>100$ Mev) and high-energy ($E>1$ Bev) protons. It is formed by two thin disks of plastic scintillator, and has a geometric factor of 13.6 cm²-ster. When a coincidence occurs, the pulse height from one of the scintillators is measured by an on-board 32-channel differential pulse height analyzer which has a storage capacity of 65,535 counts per channel. The analyzer stores information for 5 minutes and is read out for 2 minutes during each 7-minute interval.

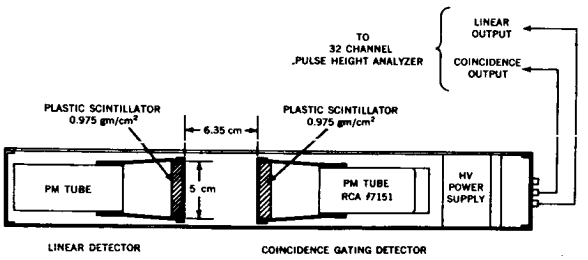


FIGURE 1.—The scintillation counter telescope. A coincidence must occur to allow the pulse height from one scintillator to be recorded. This requirement and the bias levels of the 32 channel pulse height analyzer set the lower energy limit of detectable protons at 100 Mev.

The response of the analyzer to galactic cosmic rays on 4 days is shown in figure 2 with the proton energy calibration. During these quiet times most of the particles are minimum ionizing and produce a well-defined peak between channels 7 and 8. The position of this peak serves as a calibration for all channels: A gain shift of about 4 percent between August 18 and September 6 is well marked. The distributions shown in figure 2 are caused by the Landau spread of the minimum-ionizing particles and by the higher energy losses of the lower energy particles. An overflow channel

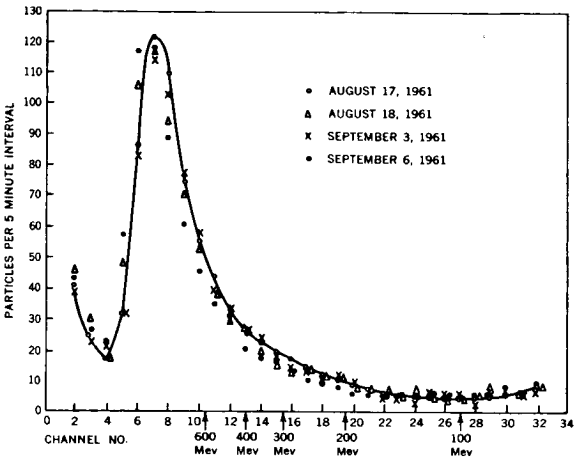


FIGURE 2.—Scintillation counter telescope pulse height distributions taken on 4 days during solar quiet times. The peak is due to minimum ionizing cosmic rays.

records large energy losses and provides a composite measurement of alpha particles and of protons between 50 and 80 Mev. Statistical fluctuations in the energy loss of alpha particles result in some of their counts appearing in the last few channels of the analyzer. Since protons of energy less than 100 Mev are counted in these channels too, the lower limit for measuring protons with the detector is set at 100 Mev. The upper limit is set at 600 Mev because all protons above this energy suffer statistical fluctuations in energy loss comparable to the difference in their energy losses. High-energy electrons can register in the minimum ionization channels but cannot do so in the 100- to 600-Mev proton range.

A thin CsI (Tl) crystal (fig. 3), 1.9 cm in diameter and 0.5 gm/cm² thick, covered only by

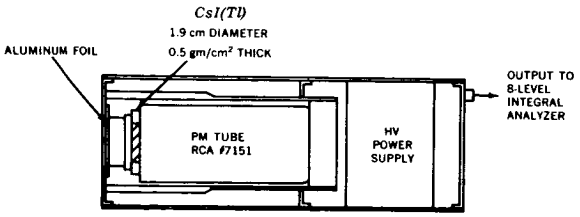


FIGURE 3.—The scintillation counter assembly. The thin cesium iodide crystal, the aluminum collimator and foil, the photomultiplier and power supply are shown. The foil and the bias levels of the 8-channel-pulse height analyzer set the lower limit of energy of detectable protons at 2 Mev.

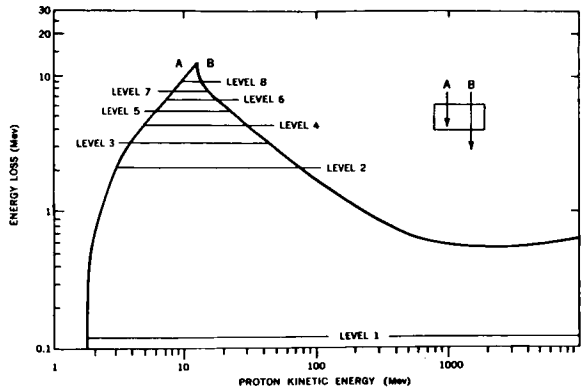


FIGURE 4.—Energy loss in the CsI crystal versus proton energy. The response is double valued since protons traversing the crystal (B) can lose the same energy as those stopping (A). Information from the other detectors is used to resolve this ambiguity.

a 6.5 mg/cm^2 aluminum foil, extended the measurements to lower energies. An aluminum collimator with an average thickness of 1.7 gm/cm^2 surrounds the crystal. The output of the photomultiplier viewing this crystal is fed to an 8-level integral pulse height analyzer. Counts are stored for 1.6 seconds once every 26 seconds at each level, with a 2.6 second interval between changing levels. Calibration is provided by a small Pu^{239} alpha particle source mounted on the front of the crystal. These alpha particles have an effective energy loss of 3.5 Mev in the CsI crystal.

The energy response for protons incident through the thin foil is shown in figure 4. The response is double valued: Group A represents stopping protons and group B represents those protons completely traversing the crystal. The single crystal data are corrected for particles in region B and for particles that are not directly incident on the foil but pass through the back and through the sides of the aluminum collimator. Assuming an isotropic distribution, data from the scintillation counter telescope and Geiger counter telescope are used to make these corrections. The 8th level, for example, then gives the differential proton intensity between 9 and 14 Mev, and levels 2 to 7 give differential intensities in the region 3 to 9 Mev. The first level is set at 120 Kev and is sensitive to electrons above 130 Kev and to protons above 2 Mev. The electron data will be discussed in a subsequent paper when bremsstrahlung, scattering, and nuclear-interaction background corrections have been evaluated.

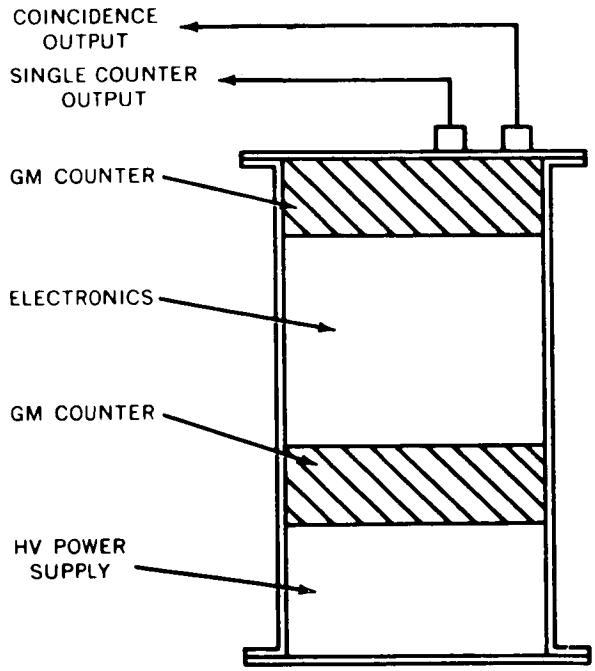


FIGURE 5.—The Geiger-Müller counter telescope. The rate of coincident events and the rate of single events in the top counter are recorded.

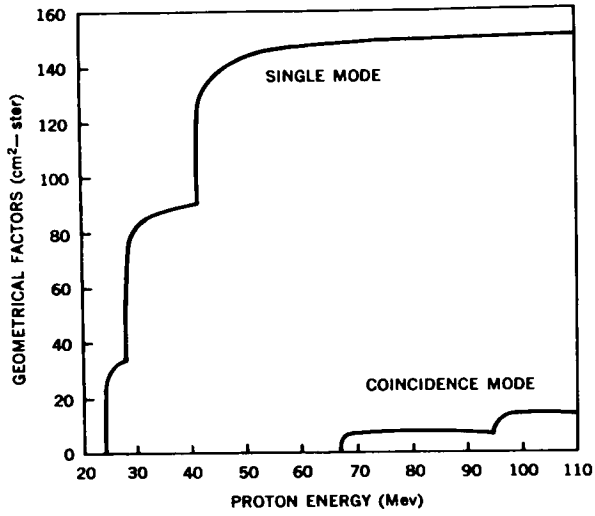


FIGURE 6.—Geiger-Müller counter geometric factors, which are defined by the shielding, are shown as a function of energy for the single and coincidence modes.

A Geiger–Müller (GM) counter telescope (fig. 5) consisting of two halogen-filled disk-shaped counters, is the third detector. The telescope counting rate and the rate of the top counter are sampled for 1.6 seconds every 26 seconds. The active volume of each counter has a diameter of 4.45 cm and a depth of 1 cm. The geometric factors of the telescope and the single counter are functions of particle energy; these are shown in figure 6. The threshold for the telescope is 70 Mev for protons and 8 Mev for electrons. Its efficiency is 88% for singly-charged minimum-ionizing particles.

By assuming an energy spectrum of the form $AE^{-\gamma}$ for solar cosmic rays, A and γ can be evaluated from the telescope and single GM counter rates. The GM counters were included primarily to assist in interpreting the scintillation counter measurements. The axis of the GM counter telescope is parallel to the satellite's spin axis and orthogonal to the axes of the scintillation counter telescope and the single CsI crystal.

The detectors were designed primarily to study protons; however, high-energy electrons could register along with minimum-ionizing protons, while low-energy electrons could be counted only on the first level of the CsI crystal analyzer. It is thus possible to construct a complete energy spectrum from 3 to 600 Mev for solar protons without electron contamination. The geometrical factors of all detectors were such that they saturated or partially saturated within the Van Allen zone.

GALACTIC COSMIC RAY INTENSITIES

Meaningful measurements of the lower energy primary cosmic radiation are more difficult than those of solar protons because of the preponderance of high energy particles at this stage of the solar cycle. These high energy cosmic rays have a greater efficiency for the production of secondaries in the 100- to 600-Mev range than do solar cosmic rays. The accuracy that can be achieved on quiet-time intensities, therefore, not only determines the extent to which we can study modulation effects but also delineates the precision that can be achieved on solar protons.

Typical pulse height distributions from the scintillation counter telescope at apogee are shown in figure 2. The primary alpha particle distribution is off-scale and cannot be resolved from very low energy protons and heavily ionizing background. A correction for the Landau distribution[2,3] is applied on the form of the curve obtained from μ mesons in a laboratory test. Only the shape of the curve is used; the location of the peak is determined from the inflight distribution. In addition, a correction is applied in the 100- to 600-Mev region for background produced by nuclear interactions. This correction, derived from balloon flights with similar detectors, is of the order of 50% in the 100- to 200-Mev region and decreases to 35% in the 400- to 500-Mev region. The differential intensity, shown in figure 7, is consistent with the data of Vogt.[4] In addition, balloon data by Bryant and McDonald[5] from Churchill at 90 Mev for protons are in agreement with the Explorer XII data. The integral intensity on August 18, 1960, of

$$Jp(Ep>600 \text{ Mev}) = 1380(\text{particles/m}^2\text{-sec-ster})$$

also agrees with that expected at this stage of the solar cycle.[6] The total counting rate of the scintillation counter telescope J_T, including off-scale counts and background, is a measure of the integral intensity of protons of energy greater than 50 Mev plus background. On August 18,

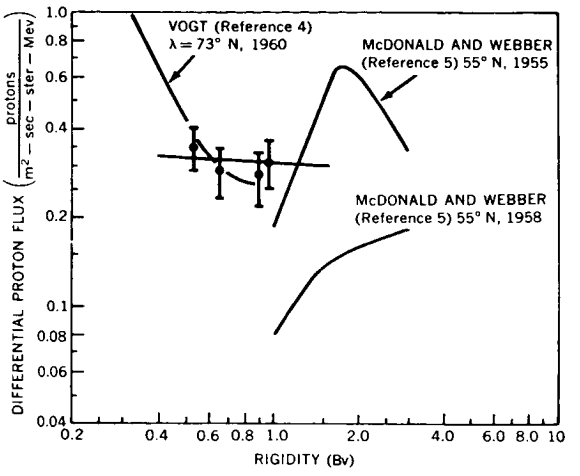


FIGURE 7.—The low energy primary cosmic ray spectrum. The differential intensity between 0.5 and 1 Bv is consistent both with the results of Vogt and with a flat spectrum at these rigidities.

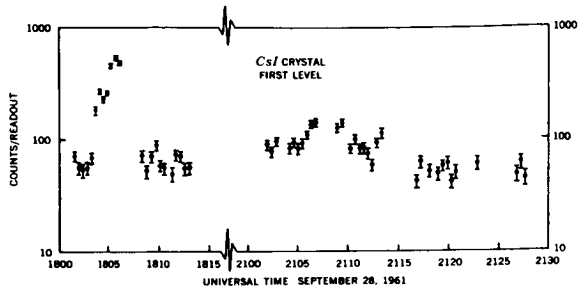


FIGURE 8.—Preflare activity. Two counting rate increases which occurred several hours before the solar flare on September 28, 1961, are interpreted as electron bursts.

1961,

$$J_T = 2100 \pm 70 \text{(particles/m}^2\text{-sec-ster)}.$$

For comparison, the total intensity measured by the GM counter telescope is

$$J_T = 1980 \pm 200 \text{(particles/m}^2\text{-sec-ster)}.$$

A detailed study of cosmic ray modulation is not made in this paper. The Forbush decrease of September 30, 1961, will be discussed in a later section.

THE SEPTEMBER 28, 1961 SOLAR COSMIC RAY EVENT

On September 28, 1961, a class 3+ solar flare occurred at 14°N and 30°E. Enhanced H_α emission began at 2202 U.T., reached a maximum at 2223, and ended at 0009 U.T. on September 29, and a type IV radio outburst occurred at 2212 U.T. (CRPL). At 2215 U.T., a short X-ray burst reached maximum (K. A. Anderson, private communication).

Preflare Activity

A few hours before the flare (at 1803 U.T. and at 2105 U.T.), when the satellite was located outside the magnetosphere at 80,000 km, two short counting rate increases were recorded by the first level of the single crystal scintillator; these are shown in figure 8. This level has a threshold of about 2 Mev for protons and about 130 Kev for electrons and gamma rays, but also responds to pileup of high-intensity low-energy electrons as observed in the outer radiation belt. The second level, having a threshold of 2 Mev for electrons and 3 Mev for protons, did not exhibit an increase. Thus, it is almost certain that these increases were caused not by protons but by either solar gamma rays or a cloud of high-energy electrons outside the radiation zone. On the basis of Explorer XII data alone we cannot decide between these two possibilities. However, Anderson, with a balloon-borne X-ray detector near the top of the atmosphere after 1800 U.T. on September 28, did not observe these increases (K. A. Anderson, private communication). We believe, therefore, that these bursts were due to electrons. The peak intensity observed in the first burst corresponds to about 120 electrons/cm²-sec-ster.

Solar Cosmic Rays

The high-energy protons which arrived soon after the flare was observed were detected by the scintillation counter telescope and the GM counter telescope. Because of noisy reception from the satellite just at the time of the flare,

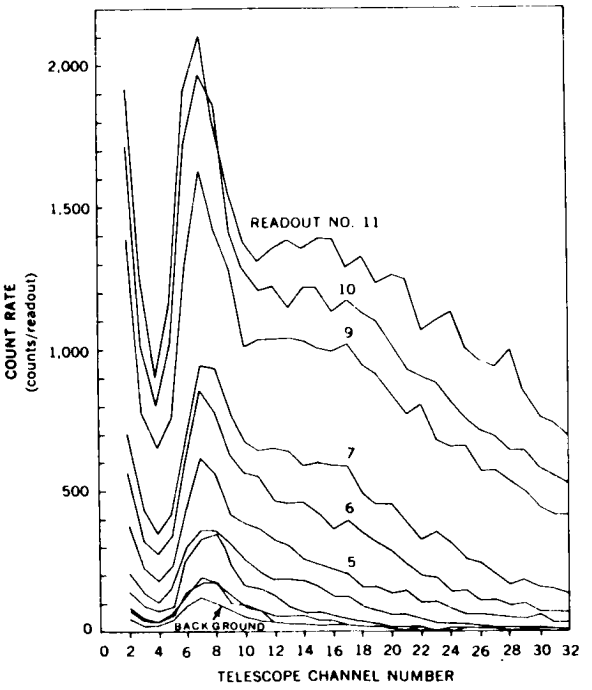


FIGURE 9.—Differential pulse height spectrum for the 32-channel analyzer taken following the solar flare on September 28, 1961. (One spectrum is missing.) The background spectrum, taken a few hours before the flare, is indicated for comparison.

reliable data were not available until after 2239 U.T. on September 28, about 24 minutes after the X-ray peak. The solar cosmic ray intensity had then reached only 50 percent of the intensity of galactic cosmic rays of energy greater than 100 Mev. Plots of the first few differential pulse height distributions recorded by the scintillation telescope after the flare are shown in figure 9, illustrating the form in which the raw data are obtained. "Background" in the figure indicates the distribution of quiet-time galactic cosmic rays. Inspection of these curves reveals that, at first, the relative increase in the intensity of minimum-ionizing particles is greater that that of the more heavily ionizing particles.

Analysis of these distributions gives the differential energy spectra shown in figure 10. The energy interval shown is between 120 and 620 Mev; the values of the integral intensities of the protons near minimum-ionization, or above 620 Mev, are not shown. It is evident that the differential intensity of the higher energy particles was greater at first but was gradually overtaken by the intensities of the successively lower energy particles. Each time-label indicates the start of the 5-minute interval during which the respective counts were stored in the analyzer. At the time corresponding to the last curve shown, the higher energy particles were reaching maximum intensity while the intensities of the particles below 100 Mev were still increasing.

Anisotropy of Medium Energy Solar Cosmic Rays

The scintillation counter telescope and the GM counter telescope were oriented on the satellite

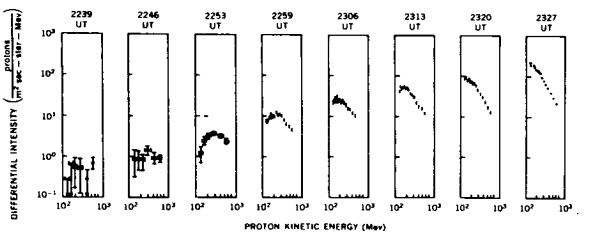


FIGURE 10.—The rise of the September 28, solar proton event, showing differential energy spectra of the particle increases over galactic background. The "snapshots" of the solar proton spectrum were recorded about 7 minutes apart and show that the spectral shape changed rapidly early in the event.

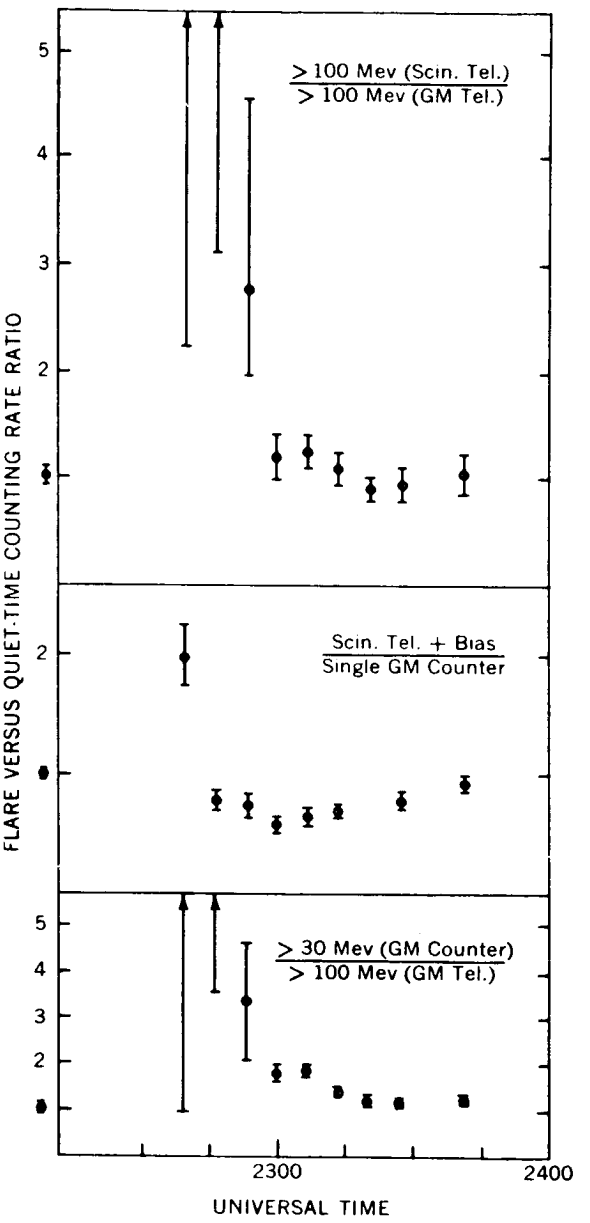


FIGURE 11.—Flare increase versus quiet-time counting rate ratios. The quiet-time isotropic values are normalized to 1. The uncertainty in the isotropic ratio is indicated by the lower bars at "1" on the ordinate scale.

with orthogonal axes; the former was perpendicular and the latter parallel to the spin axis. Since both the GM counter telescope and the scintillation telescope (used as an integral counter with the "overflow" counts omitted) had thresholds near 100 Mev, a comparison of the counting rates

of these detectors gives a direct measure of the anisotropy.

The top section of figure 11 shows the behavior of the ratio of the counting rate increases. The ratios are normalized to unity for the case of isotropy. For the first few minutes of reliable data recovery, the GM counter telescope showed almost no increase, indicating that the solar beam was essentially absent within a large-angle cone centered about the spin axis. A similar measure is provided by the ratio of the increases of the counting rates of the single GM counter and the GM counter telescope; this is indicated in the lowest section of figure 11. This comparison is valid, since at that time there were no solar particles of energy between 30 and 100 Mev.

The rate measured by the scintillation counter, including the overflow channel, and that measured by the single GM counter both give integral intensities above about 30 Mev. The ratio of increases of these rates is plotted in the center section of figure 11. An anisotropic beam should produce a ratio smaller than 1 unless the beam is confined very nearly within a plane whose normal is the spin axis, in which case the ratio should be closer to 2.5. The results, then, indicate that the solar beam began in this plane but smeared out toward isotropy within 1 hour following the flare. This trend toward isotropy of all solar protons of energy greater than 100 Mev must hold for all the energy intervals contained unless the increase in the intensity of completely isotropic medium-energy protons masked a possibly prolonged anisotropy of only the very high-energy protons.

The equatorial plane of the satellite, defined by the spin axis pointing towards 47 ± 0.5 degrees right ascension and -27.5 ± 0.5 degrees declination on September 28 (Caulk and Davis, private communication), coincided with the plane which included the direction of the solar beam early in the event. This plane intersects the ecliptic plane in a line pointing about 45 degrees west (towards negative right ascension) or 135 degrees east of the sun. Assuming that the solar protons traveled close to the plane of the ecliptic, one of these two directions is that of the apparent, or projected, proton source. The first location, 45° west, agrees with earlier determinations by McCracken,[7] which were calculated from data obtained with ground-based monitors.

Anisotropy of Low Energy Solar Cosmic Rays

The low energy solar protons also arrived anisotropically but the anisotropy was different from that observed at the medium energies. Periodic samplings of the intensity of protons at energies from 2 to 15 Mev are made by the single CsI crystal and the 8-level analyzer. Data are stored while the detector, mounted with its aperture perpendicular to the spin axis, sweeps through an angle of 290°. Since the interval between samples is not equal to the spin period, the mean direction of observation changes from one sampling to the next. A complete scan is then obtained in a number of successive readouts.

A sinusoidal modulation of particle intensity with a period equal to the predicted period showed that an anisotropy existed for many hours at low energies. For early September 29, the low counting rates make a statistical analysis necessary to reveal the anisotropy but when the intensity increases the modulation is very clear and regular. The modulation lasted at least until after the low energy solar protons reached maximum intensity between 0300 and 0500 U.T. on September 29, but had disappeared by 24 hours later on September 30. Although further analysis is required to determine the degree and direction of the anisotropy, analysis to date has shown that the measurements are inconsistent with the existence of a beam highly collimating along the "garden hose" interplanetary field lines.

Dynamics of the Solar Cosmic Rays

Following the onset of isotropy of the solar protons of energy greater than 100 Mev, successively lower-energy groups increased in intensity until about 0300 U.T. on September 29. Sample differential spectra taken throughout September 29 and 30 are shown in figure 12. The decrease in the intensity of the higher-energy components is evident; late on September 29 the slope of the proton energy spectrum from about 3 to 600 Mev was about -2, and later, on September 30, it was about -3. The maximum spectral intensity of solar cosmic rays in this

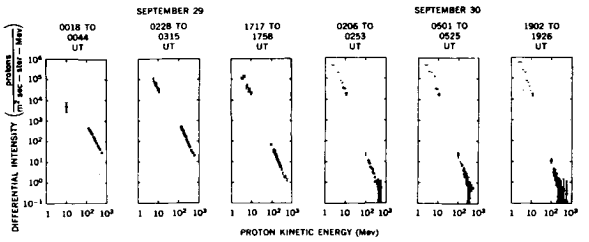


FIGURE 12.—The intensity decay of the solar proton event. These differential energy spectra, taken over two days, shown the gradual change in the medium and low energy proton intensities.

event was between 1 and 1.5 decades below that on the November 12, 1960, event.[8,9] Figure 13 shows the behavior of the *integral* intensity of $E>600$ Mev protons and the behavior of the *differential* intensities from two lower-energy groups. With decreasing energy, later onset times, steeper rises, and greater maximum intensities were observed. These features persist down to 3 Mev. The time taken for the intensity of each energy group to reach 90 percent of its maximum

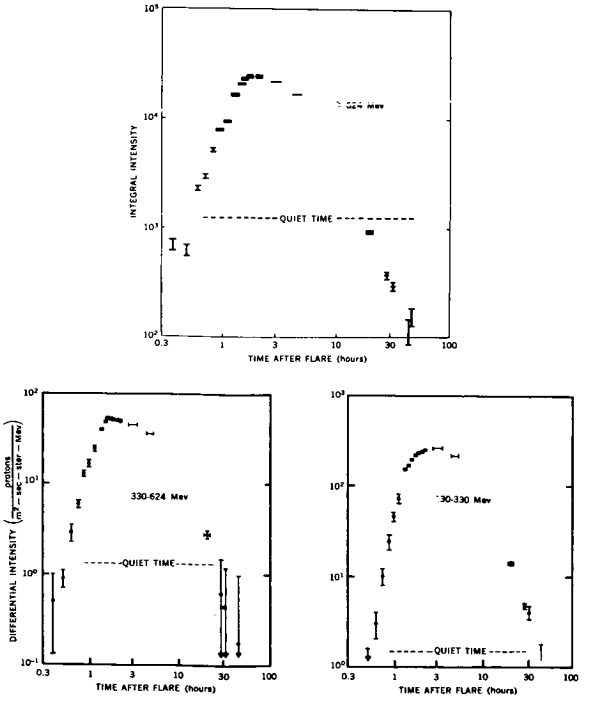


FIGURE 13.—The intensities of representative energy components of the solar protons as a function of time. Onset times are later at lower energies.

can be defined as the delay time of that spectral component. These delay times are shown in figure 14 as functions of rigidity. The rectilinear travel time versus rigidity is indicated for comparison. The dashed line indicates the duration of the anisotropy of the medium energy particles. For a large energy interval, the delay time is proportional to $R^{-1/2}$. The delay time for the higher energy protons is 8 to 9 times as long as the rectilinear travel time; for the low energy protons, it is 3 to 4 times as long.

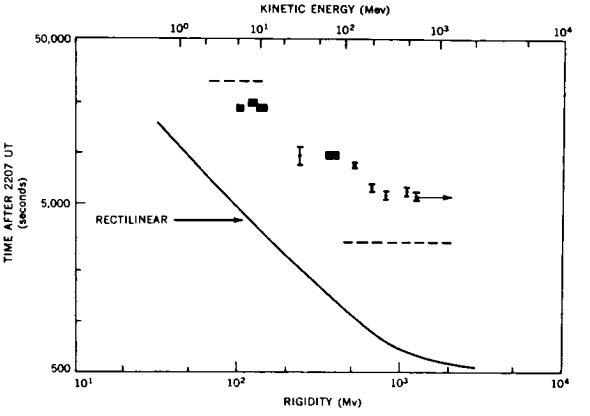


FIGURE 14.—The time to 0.9 maximum intensity of the solar protons, versus rigidity. Rectilinear travel times are indicated for comparison. The dotted lines indicate the duration of the anisotropy of the greater than 100 Mev and of the 2- to 9-Mev components.

THE ENERGETIC STORM PARTICLE EVENT OF SEPTEMBER 30, 1961

Associated with the geomagnetic storm and Forbush decrease that occurred on September 30, about 46 hours after the solar flare of September 28, was a great increase in the intensity of $E>15$ Mev protons.

Figure 15 outlines the events taking place between September 28 and October 7, showing the intensities of representative components of the radiation. The Deep River neutron monitor rate is shown for comparison and the time of occurrence of the 3+ flare on September 28 and of the sudden commencement of the magnetic storm on September 30 are indicated. On September 29 and 30 the intensities of all components of the radiation observed at the satellite were recovering, after the solar proton event of September 28,

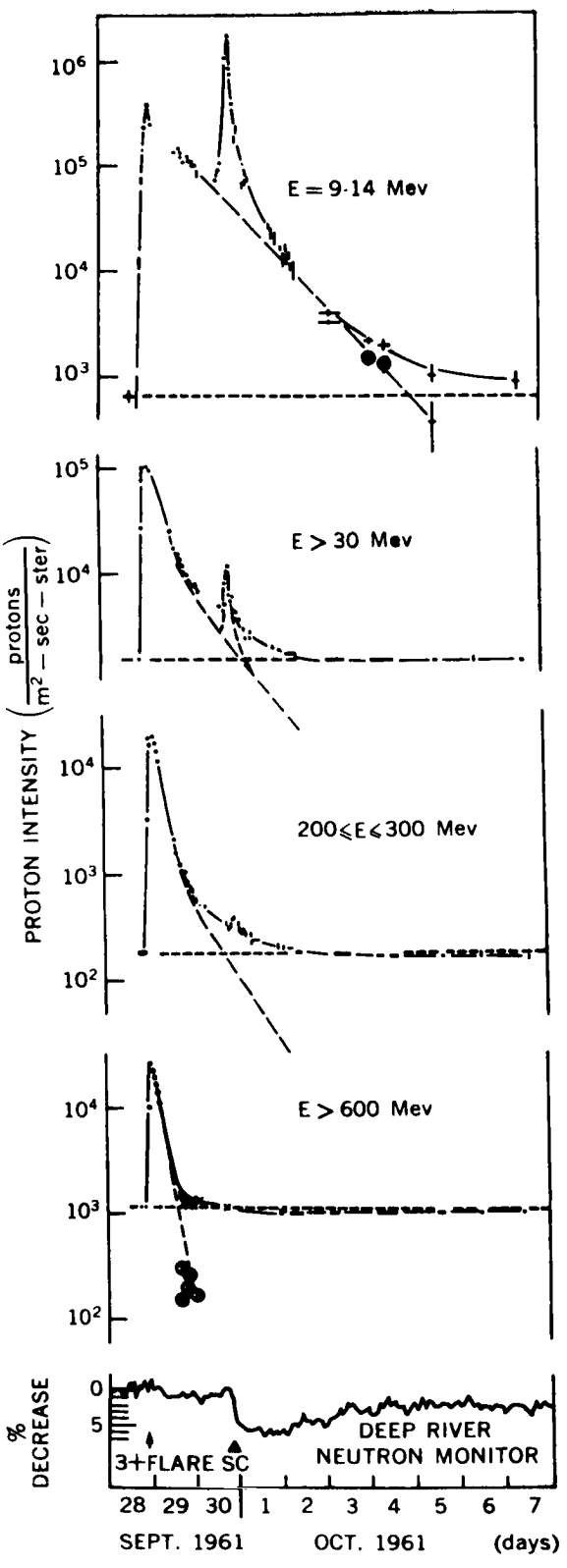


FIGURE 15.—Representative proton intensities between September 28 and October 7, 1961, showing the decay of the solar proton event and the arrival of the energetic storm particles late on September 30. The Deep River neutron monitor record is shown for comparison.

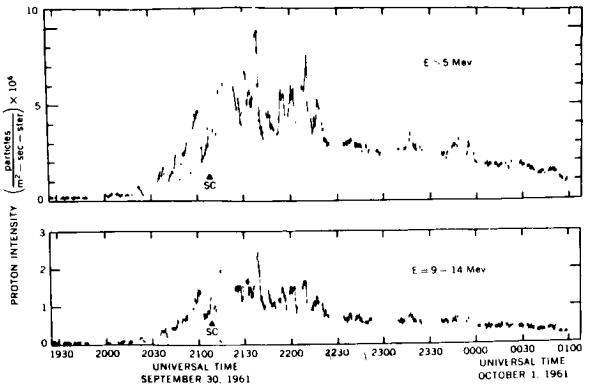


FIGURE 16.—The energetic storm particles of September 30, 1961, showing that some of the particles arrived before the sudden commencement at 2108 U.T. The intensity fluctuations are discussed in the text.

toward the preflare levels shown in the figure. The departures from preflare levels are indicated by broken lines. At about 1900 U.T. on September 30, when the satellite emerged from the radiation zone, the intensity of the greater than 600 Mev component had returned to normal and the other components were continuing their recovery. At 1930 U.T. the recovery of the low energy protons was interrupted by the arrival of the storm particles. The intensity of these particles was still rising when the Forbush decrease started and the magnetic storm began with a sudden commencement at 2108 U.T. It reached a maximum at about 2130 U.T., and had recovered to a value expected from the recovery of the solar proton event by 2000 U.T. on October 1. The recovery of the intensity of protons between 30 and 600 Mev was not obviously interrupted for more than a few hours by the passage of the plasma cloud.

Figure 16 shows the behavior of the 9 to 14 Mev component and the >5 Mev component at the time of the sudden commencement on an expanded time scale. The points are plotted, when available, at intervals of 25 seconds and represented data recorded during 1.6 second intervals. Throughout the storm particle event, both components show large fluctuations that track

very closely. Since these fluctuations are periodic and are closely connected with the spin of the satellite, they indicate a spatial anisotropy rather than a spatial or temporal variation.

Figure 17a shows energy spectra of the energetic storm particles at 2124 and 2131 U.T. on September 30. The event is essentially a low energy phenomenon and the intensities at lower energies are more than an order of magnitude greater than those reached in the solar proton event. Figure 17b shows that the low energy particles are still present on October 2 and 3 while the Forbush decrease, discussed below, is still in evidence at high energies.

THE FORBUSH DECREASE OF SEPTEMBER 30, 1961

The Forbush decrease of primary cosmic rays is shown in figure 15 by the depression of the >600 Mev component below its normal level during the first few days October 1961. Figure 18 compares the decrease observed at the satellite with that observed by the Deep River neutron monitor. On the left side of the figure the satellite data are averaged over the times indicated by the horizontal bars. The right side shows the decrease at the satellite, measured from the September 27 value, plotted against the daily mean values of the Deep River neutron monitor: the ratio is 1.7±0.3. Fan, Meyer, and Simpson,[10] comparing the decrease observed on Pioneer V (1960α) in April 1960 with the Climax neutron monitor decrease extrapolated to the top of the atmosphere, found a ratio of 1.3±0.15: our measurements in October 1961 give 0.9±0.2.

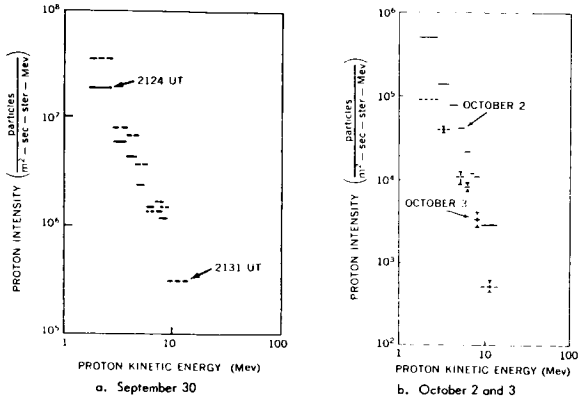


FIGURE 17.—Differential energy spectra of the energetic storm particles at 2124 and 2131 U.T. on September 30, and those on October 2 and 3, 1961.

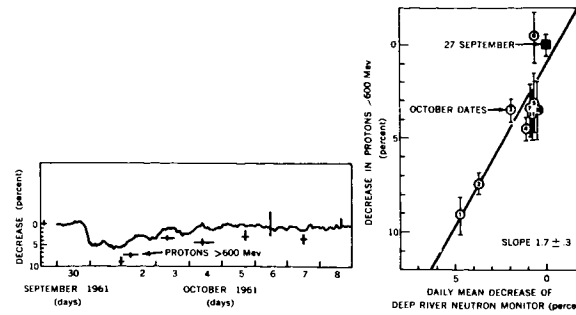


FIGURE 18.—The Forbush Decrease of September 30, 1961. The left side shows the decrease measured by the Deep River neutron monitor and the behavior of protons >600 Mev observed on Explorer XII. The right side is a plot of the decrease at Explorer XII against the decrease observed at Deep River.

DISCUSSION

There are two main magnetic configurations that have been suggested for the interplanetary medium during times of high solar activity but when no recent solar outburst has occurred: one is an essentially radial field caused by continuous emission of the solar wind and curved due to the rotation of the sun; and the other is a dipole field which produces a field perpendicular to the ecliptic at low solar latitudes.[11] Both of these configurations contain magnetic irregularities which can act as scattering centers for solar protons.

Detailed calculations have been carried out for the diffusion of solar protons caused by scattering at the magnetic irregularities when the influence of the uniform magnetic field is negligible (reference 11, and E. N. Parker, private communication). We shall confine our present discussion to this diffusion model because the quantitative calculations that have been carried out for diffusion under various boundary conditions permit the mechanism to be tested directly with Explorer XII data and the values of the parameters to be determined.

The application of classical diffusion theory yields the following expression for the time- and distance-dependence of the intensity (assumed

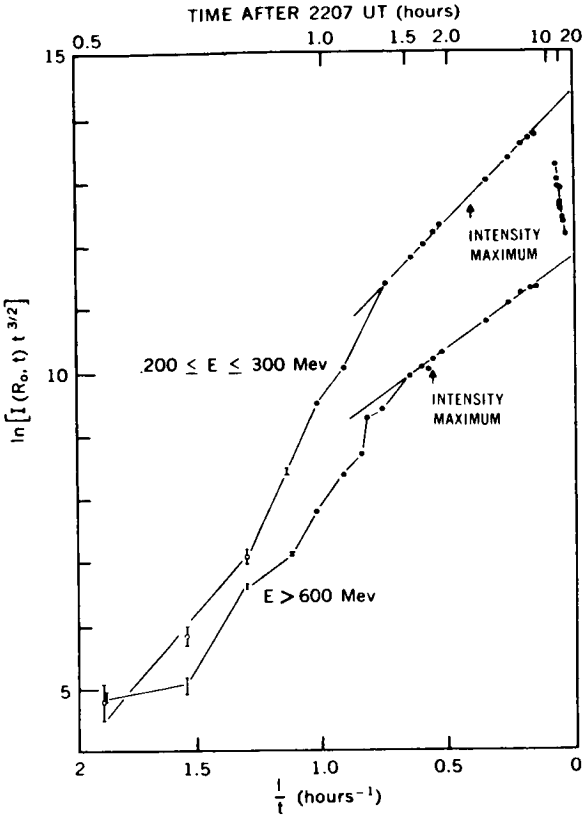


FIGURE 19.—Plot of In (Intensity $\times t^{3/2}$) against $1/t$ for protons with energies between 200 and 300 Mev and energies above 600 Mev. From 1.4 hours after the X-ray burst when isotropy is established to at least 7 hours after the X-ray burst, both components show close agreement with diffusion theory.

isotropic) of particles diffusing from a point source into an infinitely extensive medium:

$$I(R, t) = NV/32\pi[(V\lambda\pi/3)t]^{3/2} \times \exp[-\tfrac{3}{4}(R^2/V\lambda t)],$$

where

I = directional intensity,
N = number of particles released at $t=0$ and $R=0$,
R = distance from source,
t = time after release,
V = particle speed, and
λ = mean free path.

It follows that

$$\ln[I(R_0, t)t^{3/2}] = \ln[NV/32\pi(V\lambda\pi/3)^{3/2}] - \tfrac{3}{4}(R_0^2/V\lambda)t^{-1},$$

where R_0 is the astronomical unit. If the theory holds, a plot of $\ln[I(R_0, t)\, t^{3/2}]$ against $1/t$ is a straight line of slope $-3R_0^2/4V\lambda$, giving the value of λ on substitution for R_0 and V. Further, the intercept at $1/t=0$ can then be used to determine N: this aspect will not be discussed here.

Figure 19 shows $\ln[I(R_0, t)t^{3/2}]$ plotted against $1/t$ for protons >600 Mev and for protons between 200 and 300 Mev. Time is measured from the time of the X-ray burst at the sun on September 28 (K. A. Anderson, private communication). For the first 1.4 hours, when the intensity of both components was rising irregularly during the anisotropic phase of protons >100 Mev mentioned earlier, the above solution of the diffusion equation is inapplicable. From 1.4 hours, when isotropy is established, to at least 7 hours the intensities of both components show very close agreement with diffusion theory. The times of maximum intensity of both components occurred during this phase.

The slope of the straight line is −2.9 hours for protons >600 Mev, and −4 hours for protons between 200 and 300 Mev. Taking $V=0.86c$ as a typical velocity for the protons >600 Mev, and $V=0.61c$ for protons of 250 Mev,

$$\lambda/R_0 = 0.042 \text{ at } 900 \text{ Mev}$$

and

$$\lambda/R_0 = 0.043 \text{ at } 250 \text{ Mev}.$$

We see, then, that a simple diffusion model with a mean free path of 0.04 astronomical unit can explain the behavior of the solar protons near maximum intensity in this event. The fact that the mean free path is the same at values of rigidity that differ by a factor of 2 further supports the diffusion theory by indicating that large-angle scattering occurs at discrete scattering centers and that the diffusion is not strongly affected by continuous deflection in a regular magnetic field. It is clear, however, that simple diffusion in an infinite medium cannot account for the behavior late in the event, when the decay is closely exponential as is indicated in figure 15. A modification of diffusion theory, by introducing a loss term produced by a finite boundary to the medium in which the particles are diffusing (E. N. Parker, private communication), shows that the

intensity late in the event would decay exponentially. The introduction of general absorption, such as would be produced by the scattering into loss directions in the model proposed by Elliot,[12] would also introduce an exponential multiplying term into the solution of the diffusion equation. Since no information could be obtained while the satellite was in the radiation zone between 7 hours and 20 hours after the start of the event, it may not be possible to distinguish between the case of general absorption, which would introduce an exponential multiplying term, and the case of loss from a finite boundary, which would show the gradual emergence of an exponential dependence. Parker (private communication) shows that, if the diffusion takes place only within a sphere of radius P, the time constant τ of the decay late in the event is given by

$$\tau = 3P^2/\pi^2\lambda V.$$

From figure 15 we see that the time constants are about 4 hours for protons >600 Mev, and about 16 hours for protons between 200 and 300 Mev. Substitution of the values of V and λ gives $P \approx 2$ AU for protons >600 Mev and $P \approx 3$ AU for protons between 200 and 300 Mev.

In the September 28 event, then, the behavior of the solar protons can be accounted for, except during the first 1.4 hours, by diffusion within a sphere of radius 2 to 3 AU with a mean free path of 0.04 AU. This agreement with diffusion theory does not necessarily establish that the protons were diffusing within a finite medium, since no account has been taken of the influence of a general magnetic field or of the plasma cloud which was approaching the earth on September 29 and 30.

Axford and Reid,[13] have suggested a mechanism for the production of the particles responsible for the polar cap absorption that occurred immediately before the magnetic storms of February 11, 1958, and September 30, 1961. They suggest that solar protons with energies of the order of 10 Mev were accelerated by repeated reflections from the magnetic field lines bent by the shock wave ahead of the advancing plasma. Since most of the energetic storm particles seen by Explorer XII arrived after the sudden commencement, and since the plasma cloud clearly contained solar protons (fig. 15), we suggest that the particles were solar protons that were trapped in the cloud during its passage through the interplanetary medium. Turbulent flow in the plasma cloud might cause regions of enhanced magnetic field to exist. Two such regions connected by a region of weaker field would form a trap for particles whose rigidities were less than a critical rigidity P_c determined by the strength of the field and its gradient. If particles with rigidity slightly greater than P_c entered such a region and were reflected by a region of strong magnetic field moving with the plasma and away from the particle, they would loose momentum and some would be trapped. Particles of higher rigidity would be scattered. Whether solar protons could be stored in such regions during the passage of the plasma cloud to the earth depends on the way in which the cloud expands and how the magnetic fields change with time. Detailed investigations of the intensity fluctuations and directional properties of the energetic storm particles should promote further understanding of their origin.

ACKNOWLEDGMENTS

The authors are greatly indebted to Mr. A. P. Flanick for the detector development work and extensive series of laboratory tests that made these measurements possible and are very grateful to Mr. R. H. Fisk of the IBM Corporation for his valued advice and help with the processing of the data on the IBM 7090 computer.

REFERENCES

1. Desai, U. D., Van Allen, R. L., and Porreca, G., "Explorer XII Satellite Instrumentation for the Study of the Energy Spectrum of Cosmic Rays," in: *Proc. Nat. Telemetering Conf.*, Washington, vol. 1, Paper 7-5, May 1962.
2. Symon, K. R., "Fluctuations in Energy Lost by High Energy Charged Particles in Passing Through Matter," Ph.D. Thesis, Harvard University, 1948.
3. Rossi, B. B., "High-Energy Particles," New York: Prentice-Hall, 1952.
4. Vogt, R., "Primary Cosmic-Ray and Solar Protons," *Phys. Rev.*, **125**(1): 366–377, January 1, 1962.
5. Bryant, D. A., and McDonald, F. B., "*de/dx* and *E* Experiment for Primary Cosmic-Ray Measurements," *Bull. Amer. Phys. Soc.*, **7**(4): 311, April 23, 1962 (Abstract).

6. McDONALD, F. B., and WEBBER, W. R., "A Study of the Rigidity and Charge Dependence of Primary Cosmic Ray Temporal Variations," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961. *II. Joint Sessions*, Tokyo: Physical Society of Japan, pp. 428–435, 1962.
7. McCRACKEN, K. G., "The Cosmic-Ray Flare Effect," *J. Geophys. Res.*, **67**(2): 435–458, February 1962.
8. BISWAS, S., FICHTEL, C. E., and GUSS, D. E., "A Study of the Hydrogen, Helium and Heavy Nuclei in the November 12, 1960, Solar Cosmic Ray Event," submitted to *Phys. Rev.*, April 1962.
9. OGILVIE, K. W., BRYANT, D. A., and DAVIS, L. R., "Rocket Observations of Solar Protons during the November 1960 Events, 1," *J. Geophys. Res.*, **67**(3): 929–937, March 1962.
10. FAN, C. Y., MEYER, P., and SIMPSON, J. A., "Rapid Reduction of Cosmic-Radiation Intensity Measured in Interplanetary Space," *Phys. Rev. Letters*, **5**(6): 269–271, September 15, 1960.
11. MEYER, P., PARKER, E. N., and SIMPSON, J. A., "Solar Cosmic Rays of February 1956 and Their Propagation through Interplanetary Space," *Phys. Rev.*, **104**(3): 768–783, November 1, 1956.
12. ELLIOT, H., "Cosmic-Ray Intensity Variations and the Interplanetary Magnetic Field," *Phil. Mag.*, **5**(54): 601–619, June 1960.
13. AXFORD, W. I., and REID, G. C., "Polar-Cap Absorption and the Magnetic Storm of February 11, 1958," *J. Geophys. Res.*, **67**(4): 1692–1696, April 1962.

NEW EVIDENCE FOR LONG-LIVED SOLAR STREAMS IN INTERPLANETARY SPACE

D. A. BRYANT,* T. L. CLINE, U. D. DESAI AND F. B. McDONALD

Goddard Space Flight Center

Explorer XII measurements of the intensity of interplanetary protons of energy greater than 3 Mev provide a method of investigating long-lived solar plasma streams. On two occasions (October 27 and December 1, 1961) we observed an increase of the intensity of these particles. The increases were unaccompanied by a solar flare but occurred at the beginning of a magnetic storm and Forbush decrease near the time of central meridian passage of a region responsible for a flare and a solar proton event during the previous solar rotation. These increases were very small and probably could not have been detected with riometers. Chapman and Ferraro[1] have shown that a stream of neutral plasma emitted from the sun could cause magnetic storms. The recurrence of small magnetic storms over many 27-day cycles of solar rotation led Bartels[2] to postulate that such streams were continuously emitted from long-lived regions on the sun which he called M regions. Recent Mariner II measurements[3] have shown the existence of solar plasma with a 27-day structure. Magnetic fields carried by solar plasma are thought to be responsible for Forbush decreases of cosmic ray intensity which often occur during magnetic storms. Although no direct measurements of such streams have been made on Explorer XII by our detectors, the intensity increases of these >3 Mev protons required streams of plasma and magnetic fields to carry the protons from the vicinity of the sun or to accelerate them locally or to trap interplanetary solar protons remaining from a previous solar event. We feel that the presence of protons of a few Mev in the plasma does not imply that they are an intrinsic feature of the plasma but rather that it is a consequence of latent trapping regions in the plasma being filled with solar protons from a preceding event.

The instruments on Explorer XII, designed to study galactic and solar cosmic rays, have been

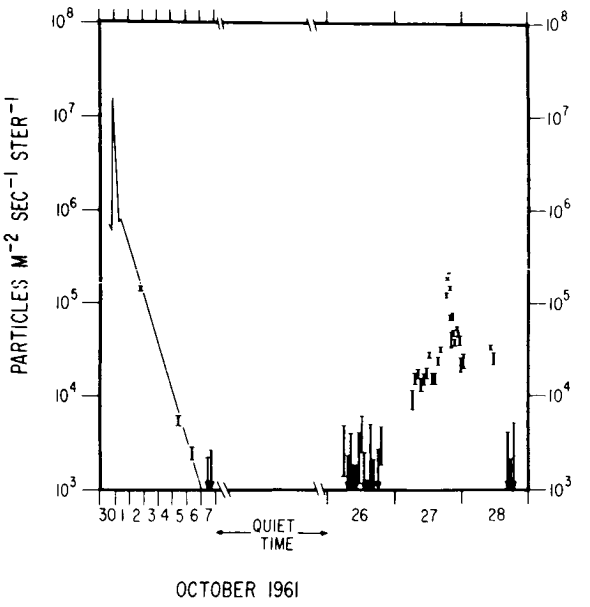


FIGURE 1.—Intensity of >3 Mev protons between September 30 and October 28, 1961 (note that the time scales before and after the period during which the intensity remained at a quiet-time value are different).

described.[4] The measurements reported here were made while the satellite was outside the magnetosphere on the sunlit side of the earth.

Figure 1 shows the intensity of interplanetary protons of energy greater than 3 Mev, whenever

*NAS–NASA resident research associate.

TABLE 1

Solar flare: (class 3 or 3+) Time: 2202 UT, 29 September 1961 Type IV radio emission: 2214 UT Location: 15°N, 29°E	Solar flare: (class >1) Time: 1435 UT, 10 November 1961 Type IV radio emission: 1440 UT Location: 19°N, ≈90°W
Forbush decrease: Duration: 30 Sept. to =5 Oct. Maximum amplitude: Explorer XII: 8≈ Deep River: 5≈	
Geomagnetic storm: Duration: 30 Sept. to ≈6 Oct. SC: 2108 UT, 30 September Main phase: 170 gamma	(No effects of two days delay)
>3 Mev protons at Explorer XII: Onset time: 1930 UT, 30 September Max. intensity: 2×10^7/m² s st	
Recurrence Phenomena Associated with Above Events	
Forbush decrease: Duration: 28 Oct. to ≈1 Nov. Maximum amplitude: Explorer XII: 3% Deep River: 2%	Forbush decrease: Duration: 1 Dec. to ≈4 Dec. Maximum amplitude: Explorer XII: 8% Deep River: 4%
Geomagnetic storms: Duration: 26 to 27 October SC: 1940 UT, 26 October Main phase: 70 gamma Duration: 28 Oct. to ≈1 Nov. SC: 0820 UT, 28 October Main phase: 280 gamma	Geomagnetic storm: Duration: 1 Dec. to ≈4 Dec. Main phase: 145 gamma
>3 Mev protons at Explorer XII: Onset time: between 26 and 27 Oct. Max. intensity: 2×10^5/m² s st	>3 Mev protons at Explorer XII: Onset time: before 0300 UT, 1 Dec. Max. intensity: $>1.6\times10^5$/m² s st

it was above the quiet-time value, from September 30 to October 28, 1961. A class 3 solar flare on September 28, 1961, initiated a solar proton event.[4] Two days later, long after the solar proton intensity had passed through a maximum, there was an increase to a level about 10 times greater than that of the solar proton maximum. At the same time there was a series of geophysical disturbances (including a magnetic storm with a sudden commencement and a mid-latitude auroral display) and a Forbush decrease of relativistic protons observed both by Explorer XII and by neutrons monitors. These events indicated the arrival of a stream of solar plasma carrying a magnetic field. The enhanced proton intensity and the geophysical disturbances had subsided by about October 7. The proton intensity remained at its normal quiet-time value from October 7 until October 27 when there was again an increase in the proton intensity accompanied by a Forbush decrease and geomagnetic storms. (Table 1 summarizes these events.)

The intensity increase on October 27 is unlike a solar proton event: it is not immediately preceded by a solar disturbance and the time constants of rise and decay are only a few hours. Further, the arrival times are not a function of proton velocity since the shapes of the differential kinetic energy spectra at onset and at maximum intensity are nearly the same, both having a power-law exponent between -4 and -5 in the range 3–10 Mev. We suggest that the active region of the sun responsible for the September 28 flare was the origin of a long-lived plasma stream that we encountered on October 27, a full rotation of the sun after September 30. If the continuous emission of this long-lived stream began during the active life of the region either on or before September 28, the >3 Mev proton intensity increase on September 30 was also caused by an encounter with that stream, but to ascribe the September 30 and October 27 proton increases to an identical phenomenon may be an oversimplification, since the earlier event may have been a result of a different and transient phenomenon (such as the shock wave postulated by Parker[5] or the magnetic bottle postulated by Gold[6]) caused by the solar activity of September 28. During the rotation of the sun following 27 October, the region stopped emitting plasma since neither our records nor geophysical observations show an event 27 days later in November.

It seems likely that our observations on October 27 are closely related to the phenomenon of *R*-rays as proposed by Müstel.[7,8] Müstel considers R-rays which are thin and filamentary extensions of the outer corona above activity centers to be responsible for recurrent magnetic storms. Figure 2 is a schematic drawing made by Müstel showing these R-rays. The fine structure in the proton intensity increase on October 27 may have been a direct consequence of the filamentary structure of the rays.

The second of the two events we wish to mention occurred on December 1, 1961. It is similar to that of October 27 and the relevant sequence of events is also outlined in table 1. On December 10, 1961, a flare accompanied by type IV emission occurred on the extreme west limb of the sun and initiated a solar proton event which was observed by Explorer XII. No magnetic storm,

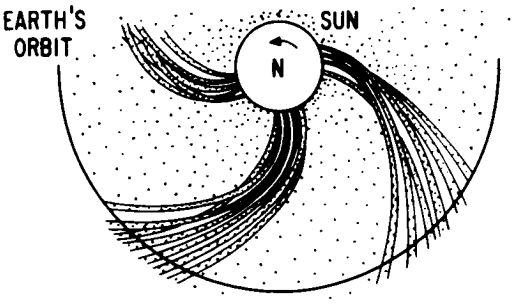


FIGURE 2.—Schematic drawing of *R*-rays after Müstel.[8]

Forbush decrease or >3 Mev proton increase was seen after the usual one- to three-day plasma transit time from the sun to the earth. Three weeks later on December 1, 1961, after three-fourths of a solar rotation, there was a Forbush decrease, a magnetic storm and a >3 Mev proton event similar to that of October 27 indicating the arrival of a solar plasma stream. The occurrence of the event of December 1 closely coincided with the central meridian passage of the active region which produced the flare on November 10. This event again substantiates the picture of a long-lived plasma stream emanating from an active region of the sun but in this case the timing was different owing to the different position of the parent flare on the sun.

There is another occurrence which could be the same phenomenon. On September 7, 1961, a solar proton event having an anomalously slow intensity decay was observed by Explorer XII, the details of which will be reported in a later paper. There was no observation at that time of a large flare or of type IV emission which can definitely be associated with this event and there was no geophysical disturbance two days later. Eleven days later on September 18 there was a small increase of the intensity of >3 Mev protons similar in spectrum to those of October 27 and December 1 and unaccompanied by a flare. We speculate from these observations that the flare responsible for the solar proton event of September 7 occurred on the remote side of the sun and that the increase on September 18 was the same phenomenon as that observed on October 27 and December 1. In contrast to this view Skerjanec et al.[9] attribute the September 7 event to a solar radio noise storm on September 6 occurring in a

plage region which was 40 degrees west on September 6 and which was the origin of a solar cosmic ray event on September 10 when this plage region reached the west limb of the sun.

REFERENCES

1. CHAPMAN, S., and FERRARO, V. C. A., *Nature*, **126,** 129 (1930).
2. BARTELS, J., *Terr. Mag. Atm. Elect.*, **37,** 48 (1932).
3. SNYDER, C. W., and NAUGEBAUER, M., Fourth International Space Science Symposium, COSPAR, Warsaw, Poland (1963).
4. BRYANT, D. A., CLINE, T. L., DESAI, U. D., and MCDONALD, F. B., *J. Geophys. Res.*, **67,** No. 13, 4983 (1962).
5. PARKER, E. N., *Space Science Reviews*, **1,** 62 (1962).
6. GOLD, T., *J. Geophys. Res.*, **64,** 1665 (1959).
7. MUSTEL, E. N., *Astronomicheskii Zhurnal*, **39,** No. 3, 418 (1962), *Soviet Astronomy*, **6,** No. 3, 333 (1962).
8. MUSTEL, E. N., *Astronomicheskii Zhurnal*, **39,** No. 4, 619 (1962), *Soviet Astronomy*, **6,** No. 4, 488 (1963).
9. SKERJANEC, R. E., WHITEMAN, D. W., WARWICK, J. W., Information Bulletin of Solar Radio Observatories, **13,** 5 (Feb. 1963).

AUTOMATIC MAPPING OF THE GEOMAGNETIC FIELD

JOSEPH C. CAIN

Goddard Space Flight Center

JAMES R. NEILON

U.S. Weather Bureau

Spheric harmonic representations are used to illustrate the application of the U.S. Weather Bureau machine-mapping system for drawing charts of the geomagnetic field. This technique is also applied to the representation of model S_q fluctuations by maps of electric current flow in the ionosphere.

INTRODUCTION

The purpose of this note is to discuss the application of the field-contouring techniques developed at the U. S. Weather Bureau to the illustration of the geomagnetic field and its time changes. This study, an outgrowth of the work with Vanguard 3 (1959η) magnetic data (Cain et al., 1962) and the analysis of magnetic survey data (Jensen and Cain, 1962), was made in connection with preparation for analysis of the magnetic data to be obtained with the polar orbiting geophysical observatory (Ludwig, 1963).

The need for a general system of rapid illustration in geomagnetism has arisen with the advent of methods allowing automatic processing of large quantities of data. The techniques have been found of value not only for illustrating the results themselves but also in investigating the various steps of analyses.

CONTOURING PROCEDURE

The procedure followed (Bedient and Neilon, 1962) is first to produce coarse grid point values of the function to be contoured and then to perform a two-dimensional quadratic interpolation to locate the contour lines on a much finer grid. In the application illustrated here the original field is computed on a $\frac{1}{2}$-inch grid 39 units in latitude and 55 units in longitude, a total of 2145 points. The interpolation is then performed by means of Bessel's central difference interpolation formula (second degree) to produce a further division of this interval by a factor of 8 so that

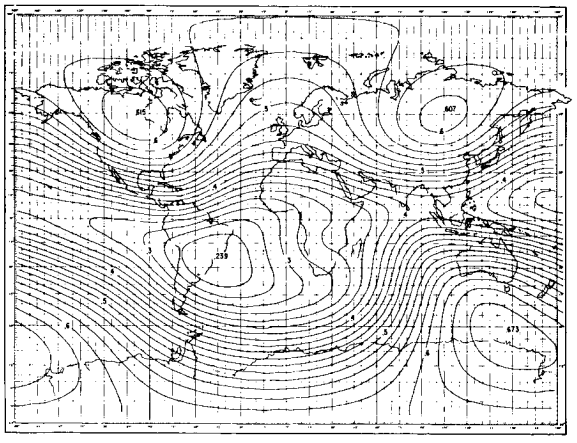


FIGURE 1.—Contours of the total intensity of the geomagnetic field in gauss, synthesized from a set of 48 spherical harmonic coefficients (Jensen and Cain, 1962) for epoch of 1960.0 (contour interval 0.02 gauss). All centers are highs except the 0.239 over South America.

the 'decision field' is a two-dimensional $\frac{1}{16}$-inch grid. The decision field is analyzed, and the plotter pen instructions are generated on a $\frac{1}{8}$-inch square grid for drawing contours between the 'high'

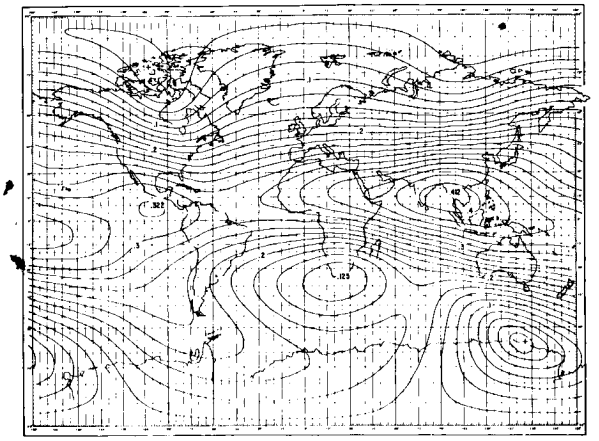


FIGURE 2.—Contours of horizontal intensity from same spherical harmonic expansion as figure 1 (contour interval 0.02 gauss). The plus signs are the positions of the north and south geomagnetic poles using this expansion. The two centers near the Equator are highs; the center near the tip of South Africa is a low.

and 'low' points in the field. The available variable control features include the selection of a particular contour and the contour interval to be used. The contour programs include sufficient corrections to compensate for the dynamics of the plotter system with a specified electronic filter network.

1960.0 MAGNETIC FIELD

The magnetic field illustrated here is that computed by Jensen and Cain (1962) for the epoch of 1960.0 using a selected set of 7400 magnetic

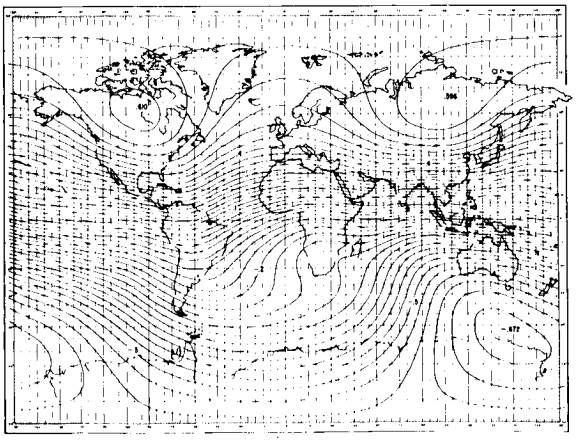


FIGURE 3.—Contours of vertical intensity from the same spherical harmonic expansion as figure 1 (contour interval 0.03 gauss). Positive highs appear in Canada and Russia, and a negative high in Antarctica.

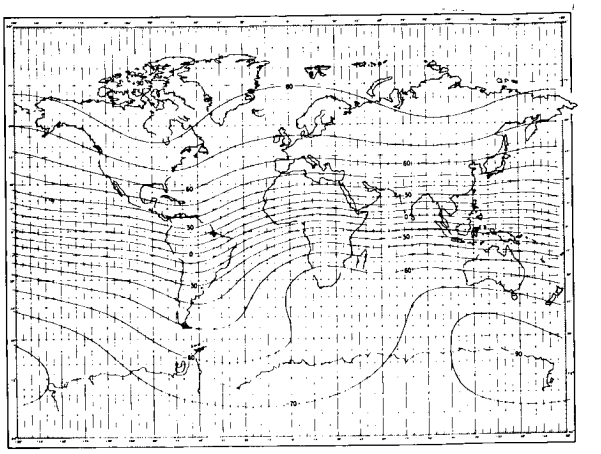


FIGURE 4.—Contours of inclination from same spherical harmonic expansion as figure 1 (contour interval 10°).

field observations for the period 1940–1961. Figures 1, 2, 3, and 4 represent the contours of F, H, Z, and I, respectively, from this 48-term spherical harmonic expansion. [The map projection is a modification of the Miller Cylindrical (Miller, 1942) as published by Army Map Service.] The three-digit numbers printed in the open areas are the highs and lows in the fields positioned approximately at the decimal point of the numbers. The large-scale kinks that usually appear in manually contoured magnetic maps are significantly absent from these four figures. How-

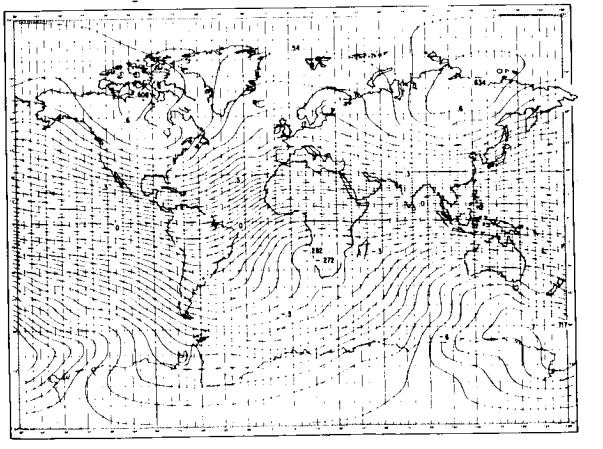


FIGURE 5.—Contours of vertical intensity synthesized from the set of 512 spherical harmonic coefficients (Jensen and Whitaker, 1960) for epoch 1955.0 (contour interval 0.03 gauss). Major highs in Canada, Russia, and Antarctica are nearly the same as those in figure 3. In addition, a −0.282-gauss high and a −0.272 low appear in Africa.

ever, there are slight waves in a few of the lines, owing to the $\frac{1}{8}$-inch grid spacing and to the dynamics of the plotting system. These small irregularities vary for different plotting equipment. Tests have shown that, on the map scales illustrated, the contours are accurate to about $\pm 1°$ of latitude. The pen speed for this system is about 20 cm/sec, so that the average plotting time is of the order of 3 minutes for each map.

To compare this relatively smoothed version of the field with a more complex map, in figure 5 the Z-component contours have been plotted from the 512-term expansion of Jensen and Whitaker (1960). Comparing the chart of this higher-order expansion with figure 3 shows that it is possible with this plotting system to illustrate much more detail than that contained in a 48-term expansion. That figure 5 is not as detailed as the original 1955 Z charts (USN Hydrographic Office), from which the data for the original analysis were taken, is due primarily to the fact that this analysis utilized data at only 10° intervals in latitude and longitude. This additional smoothing could not have occurred in the contouring, since the plotting system utilized a grid of 2145 points, whereas the original analysis by Jensen and Whitaker only employed approximately 614 observed values of Z. With this particular grid system it should be possible to represent spherical harmonic expansions of the field containing more than the 512 coefficients used. Of course, the possible map detail is limited only by the input data and scale of the map being used, so that any degree of detail can be represented by suitably expanding the scale and plotting the map by sections.

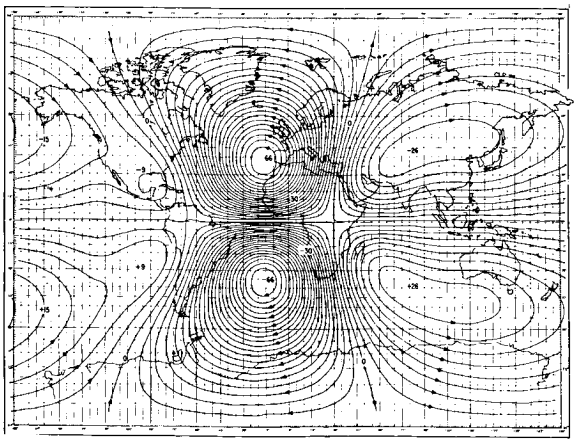


FIGURE 6.—Model S_q ionospheric current system with the sun on the Greenwich meridian (1200 U.T.). Values of current flow are in kiloamperes with a contour interval of 3000 amp. The current flows in the direction of the arrows with no current flowing across the equator. The spherical harmonic expansion is taken from Chapman and Bartels (1940) for the mean equinox of 1902, assuming a current height of 100 km.

MAGNETIC VARIATIONS

This technique of automatic plotting has also been investigated for the illustration of the time variations in the field as well as the main field itself. This application is illustrated here only by the representation of the quiet solar daily variations S_q. Figure 6 plots an idealized ionospheric current system computed by M. Sugiura (private communication) from a set of harmonic coefficients (Chapman and Bartels, 1940) representing the averaged quiet field variations at the equinoxes. Although this particular representation was produced from analytic functions, it is likely that the system will be adaptable to the illustration of worldwide magnetic variations on a synoptic basis either by representative ionospheric current stream functions or simply by contouring changes in specific components as abstracted from magnetic variometer records.

ACKNOWLEDGMENT

We wish to acknowledge the able assistance of S. Hendricks in preparing the diagrams.

REFERENCES

BEDIENT, H. A., and NEILON, J. R., Automatic production of meteorological contour charts, *Conference on Data Handling, Reduction, and Interpretation in Geophysics, Yorktown Heights, New York*, 1962.

CAIN, J. C., SHAPIRO, I. R., STOLARIK, J. D., and HEPPNER, J. P., Vanguard 3 magnetic field observations, *J. Geophys. Res.*, **67**, 5055–5069, 1962.

CHAPMAN, S., and BARTELS, J., *Geomagnetism*, vol. 2, p. 692, Table 4, Clarendon Press, Oxford, 1940.

JENSEN, D. C., and CAIN, J. C., An interim geomagnetic field (abstract), *J. Geophys. Res.*, **67**, 3568–3569, 1962.

———, and WHITAKER, W. A., A spherical harmonic analysis of the geomagnetic field (abstract), *J. Geophys. Res.*, **65**, 2500, 1920.

LUDWIG, G. H., The orbiting geophysical observatories, *Space Sci. Rev.*, in press, 1963.

MILLER, O. M., Notes on cylindrical world map projections, *Geograph. Rev.*, **32**, 424, 1942.

NEGATIVE CORRELATION BETWEEN POLAR-CAP VISUAL AURORA AND MAGNETIC ACTIVITY

T. N. DAVIS*

Geophysical Institute, University of Alaska

The relationship between the local (K) and planetary (K_p) magnetic activity and the occurrence of visual auroras inside the auroral zone is investigated. The visual aurora above geomagnetic latitude 80° is found to be negatively correlated with the K and K_p magnetic indexes, whereas nearer to the auroral zone the correlation is positive. The relationship in the region 75–80° latitude is more complicated and of a transitional nature.

INTRODUCTION

A number of investigations, including those by Gartlein (1944), Heppner (1954), and Davis (1962b), have shown a close association between the occurrence of visual aurora and local magnetic disturbance at and outside the auroral zone. On the other hand, others, particularly Feldstein (1959) and Akasofu (1963), have indicated that no association or an inverse relationship may exist between the visual aurora and magnetic disturbance well inside the auroral zone. This paper reports a further investigation of the degree of association between the occurrence of visual auroral forms and magnetic activity inside the northern auroral zone.

Relationships Between Visual Aurora and Magnetic Disturbance Near the Auroral Zone

From a study of the alignments and apparent motions of visual auroras and of the direction of the horizontal component of local magnetic disturbance, Davis and Kimball (1962) found a strong tendency for perpendicularity between the direction of horizontal disturbance and the direction of alignment of nearby auroral forms. Further, it was found that the direction of overhead ionospheric current causing the local magnetic disturbance was usually opposite to that of the observed auroral motion. Such observations and others previously cited suggest a rather close association between the visual aurora and ionospheric current at the auroral zone.

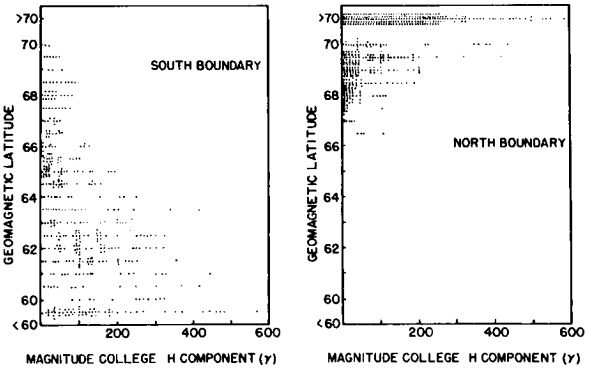


FIGURE 1.—Position of north and south boundaries of auroral displays over Alaska as a function of magnetic disturbance during the displays of February 12–26, 1958.

Diagrams given by Davis (1962a, figures 12 and 13) indicate that the occurrence or incidence of visual aurora at the auroral zone increases with baylike disturbances of the horizontal component such that the greatest incidence tends to remain at the auroral zone irrespective of the strength or direction of the auroral electrojet. However, as the local magnetic disturbance increases, the aurora becomes more distributed in latitude; both

*Now National Academy of Sciences—NASA resident research associate at Goddard Space Flight Center, Greenbelt, Maryland.

the north and south limits of the display move in opposite directions away from the auroral zone as is shown in figure 1. Akasofu and Chapman (1962) have shown that, during intervals of magnetic storms in which Dst is large, the region where the aurora occurs moves equatorward; that is, both the northern and the southern limits of the display shift equatorward. Thus it appears that near the auroral zone two separate relationships exist between the spatial distribution of the visual aurora and magnetic disturbance: (1) the meridional extent of the display becomes larger concurrently with increase in the magnitude of the auroral electrojet, and (2) the position of the latitudinal strip along which the auroral display occurs shifts equatorward as Dst becomes very large.

Relationships Between the Visual Aurora and Magnetic Disturbance Inside the Auroral Zone

Among the questions that may be asked about the polar-cap aurora are: (1) What is the relationship between the visual aurora and the local magnetic activity? (2) What is the relationship of the visual aurora to global magnetic activity? Before considering these questions we may discuss our measures of auroral and magnetic activity.

The index of visual aurora activity used here is one devised to utilize direct scalings of all-sky camera films and published Ascaplots (*Annals of IGY*, vol. 20, part 1), also derived from all-sky films. This index is a form of percentage hourly occurrence in which the hours when auroral forms are observed are weighted according to their latitudinal extent of sky coverage. Observations are made only over the region within 200 km of each observing location. Only dark periods with satisfactory observing conditions are included in compiling the auroral index. The index refers to the occurrence of distinct visible auroral forms as distinguished from widespread glows, which because of their uniformity or low brightness are not readily detected on the all-sky film. The numerical values of the auroral index given in the following text and diagrams are approximately equal to percentage values; that is, they indicate approximately the percentage of observing hours during which aurora was seen over the station.

TABLE 1.—*The Distribution of Thule K-Index Values at the Various Levels of Planetary Activity K_p for the Period January 1 to June 30, 1958*

K_p	K_{TH}									
	0	1	2	3	4	5	6	7	8	9
0		6	7	9	5					
1		14	53	52	60	34	8	1		
2		1	37	102	77	43	10	1		
3			10	70	137	105	34	7		
4			2	39	148	124	32	7		
5					36	64	28	5		
6					7	32	10	5	2	
7					1	6	1	3		
8						3	5	1	1	
9								1		1

No attempts should be made to relate these values to the particle fluxes causing the aurora.

An index of local magnetic activity is the 3-hour index K_k, the magnitude of which is related to the range of the maximum magnetic variation during a 3-hour interval at the particular observing station k. The 3-hour planetary magnetic index K_p is compiled from the K_k values obtained from a distribution of individual stations. Since the K-index scales are nonlinear, it is convenient for some purposes to convert to the linear amplitude scales a_k or a_p (Bartels, Romana, and Veldkamp, 1962). In making use of these measures of magnetic activity it must be remembered that they refer to the range of magnetic field variation during a given time interval and do not necessarily describe the total distortion to the quiet geomagnetic field.

It is of interest to examine the relationship between the local magnetic activity well inside the auroral zone with the planetary activity indexes K_p and a_p. For this purpose three stations are chosen: Thule at geomagnetic latitude 88°N is nearly at the geomagnetic pole; Godhaven at 80°N is approximately midway between the pole and the auroral zone; and College, geomagnetic latitude $64\frac{1}{2}$°, is just outside the auroral zone. Tables 1 and 2 show the relationship between the local K indexes and the planetary indexes for Thule and Godhaven, respectively. The same data

TABLE 2.—*The Distribution of Godhavn K-Index Values at the Various Levels of Planetary Activity K_p for the Period of January 1 to June 30, 1958*

K_p	K_{GO}									
	0	1	2	3	4	5	6	7	8	9
0		9	15	3						
1		30	68	77	40	9				
2		6	83	108	57	21	1			
3			41	160	105	46	22	1		
4			8	109	131	72	25	1		
5				18	55	41	15	2		
6				2	12	23	6	2		1
7				1	4	4	2			
8				1	1	5	2			1
9						1				

and similar data from College are converted to linear amplitude values a_k and a_p to allow forming of averages; the results are plotted in figure 2. It is evident from the tables and figure 2 that the local K indexes and the derived average amplitude values a_k increases with increasing planetary activity. The rate of increase is greater at the auroral zone than inside; for each of the stations College, Godhaven, and Thule, however, the average amplitude $\bar{a}_k$ increases at a lesser rate than the planetary index a_p.

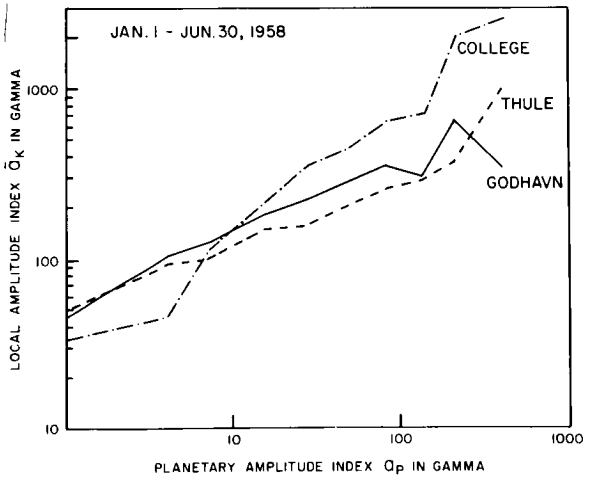


FIGURE 2.—Plots of 3-hour average amplitude indexes $\bar{a}_k$ vs. the 3-hour planetary amplitude index a_p; log scale.

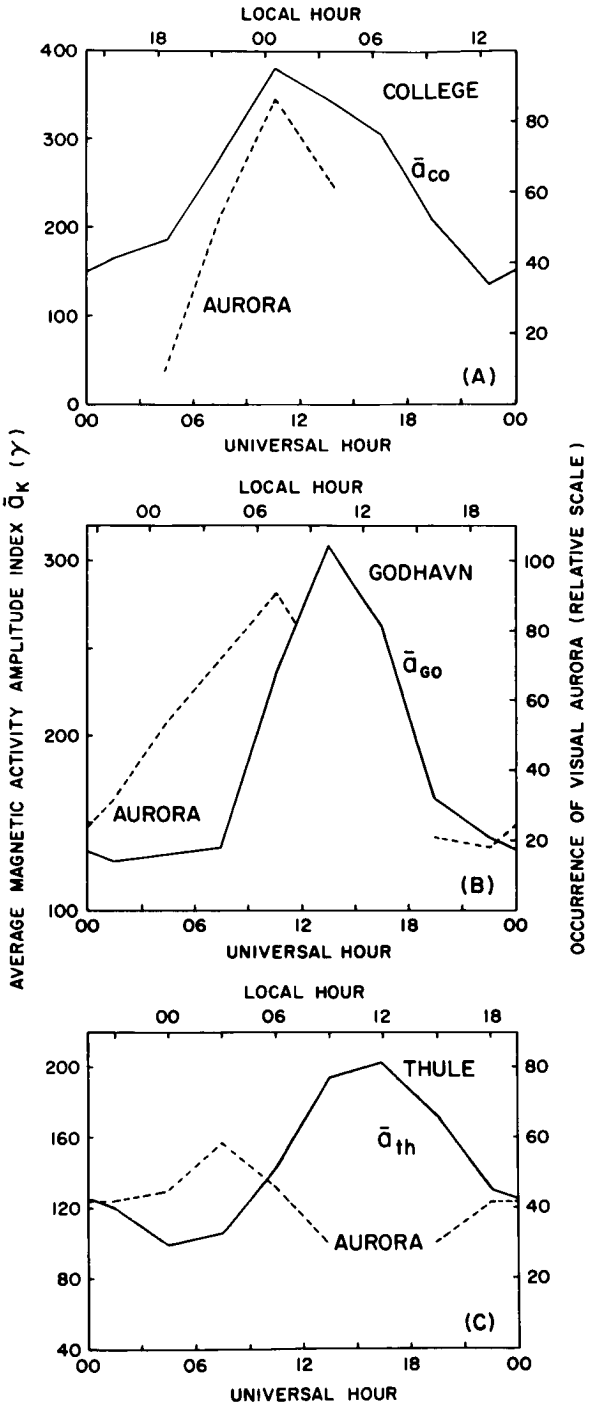


FIGURE 3.—The diurnal variation of the 3-hour amplitude index of planetary activity a_k and of the occurrence of visual aurora at (A) College, (B) Godhavn, and (C) Thule.

The diurnal variations of the average linear amplitude indexes $\bar{a}_k$ at College, Godhaven, and Thule are shown in figure 3, together with similar plots of the auroral occurrence index at each station. All auroral data used here resulted from the 1957–1958 observing season except that 1958–1959 observations are included with the Thule data to build up a better sample distribution for that station. Magnetic data from January 1 to June 30, 1958, a period of high magnetic and auroral activity, were used to compile figure 3.

From figure 3A it is seen that the maximums of the auroral and local magnetic activity occur at the same time at College, near the auroral zone. Just the opposite happens at Thule, near the geomagnetic pole, the two diurnal variation curves for that station (figure 3C) being out of phase by almost one-half day. At a station midway between the auroral zone and the pole, Godhaven (figure 3B), much of the observed auroral activity occurs during a period of the day when the local magnetic activity is low, but then the auroral activity increases with increasing magnetic activity and reaches a peak a few hours before the peak in magnetic activity.

The relationship of the visual auroral occurrence at these three stations to the planetary

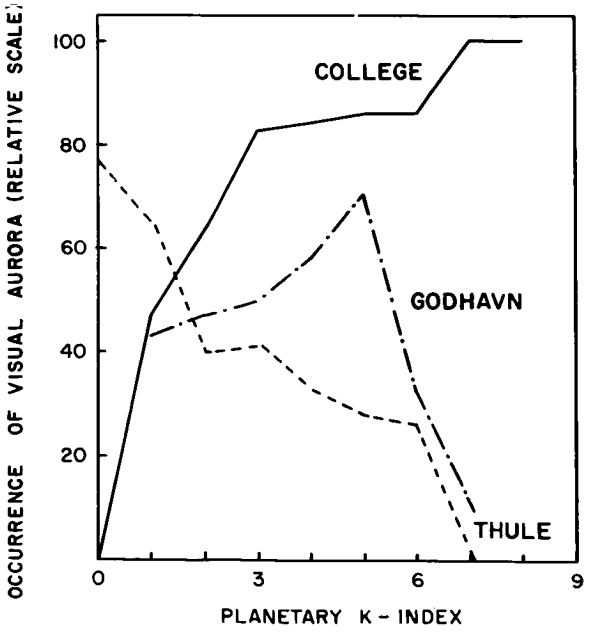


FIGURE 4.—The occurrence of visual aurora plotted as a function of the planetary magnetic activity index K_p.

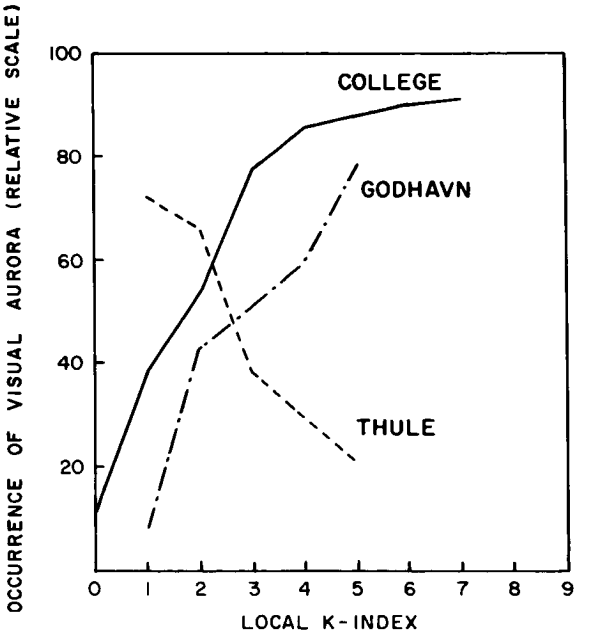


FIGURE 5.—The occurrence of visual aurora plotted as a function of the local magnetic activity index K_k.

index K_p is shown in figure 4 by using data from the same period as in figure 3. As is expected, figure 4 shows that the visual auroral activity at College, near the auroral zone, increases with increasing K_p. On the other hand, Thule, located near the geomagnetic pole, shows a decrease in the occurrence of visual auroras as K_p increases. Godhaven, midway between the pole and the auroral zone, shows a more complex relationship.

The auroral data of figure 4 are plotted again in figure 5, now against the local K indexes. Except for a more limited range of K values for which sufficient data are available, the curves of figure 5 are nearly identical with those of figure 4, and so the above discussion of figure 4 can be applied to figure 5 by replacing K_p with K_k. The dependence upon local time of the aurora and local magnetic activity at Thule could create by itself the negative correlation observed at that station. This possibility was investigated by dividing the observations made during 0600–0900 U.T. into two groups according to the level of local magnetic activity. The auroral index for the hours when $K_{\text{Thule}}=0$, 1, 2 was 82; an auroral index value of 58 occurred for the group in which $K_{\text{Thule}}=3$, 4, 5. Thus the negative correlation

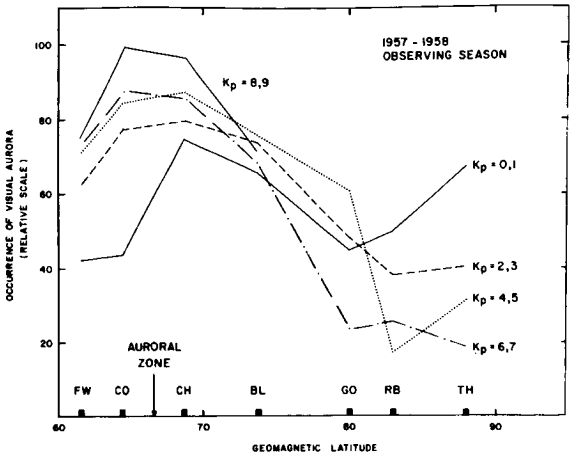


FIGURE 6.—The occurrence of visual aurora plotted as a function of geomagnetic latitude for various levels of planetary magnetic activity. Observations come from Farewell, Alaska (FW); College, Alaska (CO); Churchill, Canada (CH); Baker Lake, Canada (BL); Godhavn, Greenland (GO); Resolute Bay, Canada (RB); and Thule, Greenland (TH).

between aurora and local magnetic activity at Thule is independent of a local time effect.

More complete information about the occurrence of visual aurora as a function of geomagnetic latitude and planetary magnetic activity is presented in figures 6 and 7. The occurrence of visual aurora at Baker Lake, see figure 7, is relatively constant at all levels of magnetic activity. Below the latitude of Baker Lake a positive correlation exists between the visual aurora and the planetary magnetic activity. At latitudes above 80°N a negative correlation exists between the visual aurora and the magnetic activity.

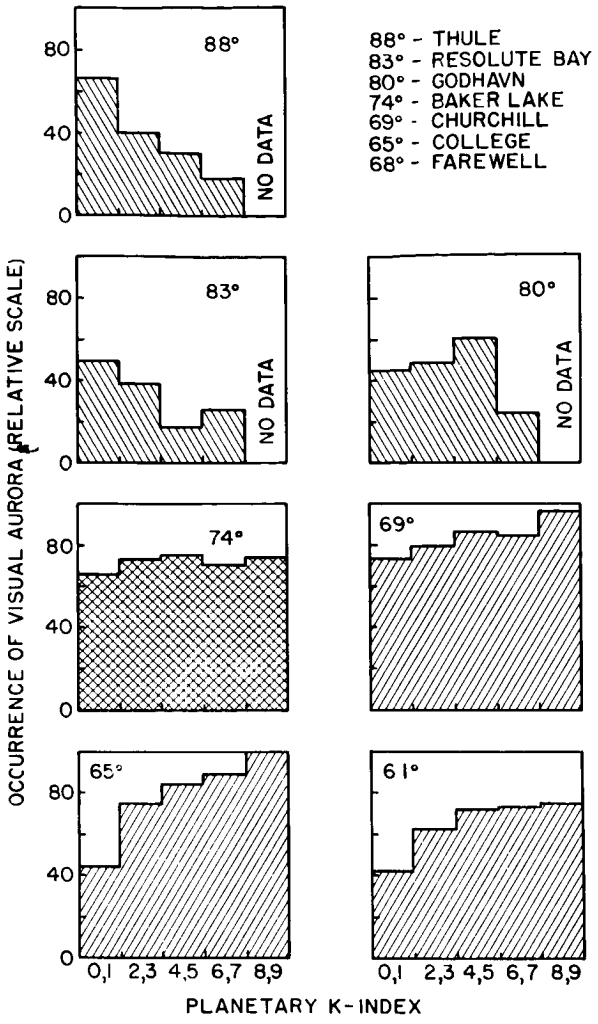


FIGURE 7.—Levels of occurrence of visual aurora versus planetary magnetic activity at seven latitude positions. Observations are from the stations listed in the upper right.

DISCUSSION AND CONCLUSIONS

The data presented here show conclusively that at very high geomagnetic latitudes the occurrence of structured visual aurora decreases as local and planetary magnetic activity increases. The demarcation between this polar-cap region and the lower-latitude region, where the occurrence of visual aurora correlates positively with the local and planetary magnetic activity, occurs at 75–80° geomagnetic latitude.

This 75–80° region also serves as an approximate boundary across which appear several differences in the morphological features of the visual aurora: (1) The polar-cap auroras are comparatively weak and short-lived. (2) They tend to exhibit more ray structure than auroras occurring near the auroral zone. (3) The auroras at the auroral zone tend to be aligned in the east-west direction, whereas the polar-cap auroras vary in the orientation so that the average orientation is in the direction approximately along the 1000–2002 meridian. (4) Motions of irregularities along auroral forms are pronounced

at the auroral zone but much less obvious in the polar-cap region. (5) The diurnal behavior of the polar-cap aurora differs from that at the auroral zone.

These morphological differences and the opposite correlation between visual aurora and magnetic activity across the 75–80° latitude boundary suggest the possibility of a difference in the origin of the auroras occurring on the two sides of the boundary. A projection of the 75–80° boundary along a dipole geomagnetic field falls at a distance of 15 to 35 earth radii. Results from Explorer 10 (Heppner et al., 1963) indicate that the dark-side magnetosphere boundary occurs in or beyond this distance range. If the projection of the 75–80° boundary coincides with the magnetosphere boundary near or above the equatorial plane, the polar-cap auroras may result from particles penetrating the magnetosphere away from the equatorial plane and passing directly to the polar atmosphere. During times when particles cross the magnetosphere boundary, it may be necessary for the local geomagnetic field to have a component normal to the boundary and thus not be completely confined by it.

An alternative interpretation of the observations involves the supposition that both polar-cap and auroral-zone visual auroras are caused by completely trapped primary particles. The paths of the primaries causing the polar-cap visual aurora are then largely in regions of low magnetic field strength and therefore are easily affected by transients arising from variation in strength of the solar wind. This factor could explain the negative correlation between the occurrence of the polar-cap visual forms and magnetic activity.

ACKNOWLEDGMENTS

I am grateful to Dr. Masahisa Sugiura for helpful discussions during the preparation of this paper.

This research has been supported by the National Science Foundation, grant NSF-G14782.

REFERENCES

Akasofu, S.-I., The dynamical morphology of the aurora polaris, *J. Geophys. Res.*, **68,** 1667–1674, 1963.

———, and Chapman, S., Large-scale auroral motions and polar magnetic disturbances, 3, *J. Atmospheric Terrest. Phys.*, **24,** 785–796, 1962.

Bartels, J., Romana, A., and Veldkamp, J., Geomagnetic data 1958, indices *K* and *C*, *ISGA Bull.* 12 *m* 1, 112 pp., 1962.

Davis, T. N., The morphology of the auroral displays of 1957–1958, 1, Statistical analyses of Alaska data, *J. Geophys. Res.*, **67,** 59–74, 1962*a*.

———, The morphology of the auroral displays of 1957–1958, 2, Detail analyses of Alaska data and analyses of high-latitude data, *J. Geophys. Res.*, **67,** 75–110, 1962*b*.

———, and Kimball, D. S., The auroral display of February 13–14, 1958, *Rept. UAG–120*, Geophysical Institute, College, Alaska, 24 pp., 1962.

Feldstein, Y. A., Distribution of aurorae in the circumpolar region, translation, Information on Soviet Bloc International Cooperation—1959, U.S. Department of Commerce, Office of Technical Services, Washington, D.C., 1959.

Gartlein, C. W., Relation of three-hour-range indices *K* to aurora seen at Ithaca, New York, *Trans. Am. Geophys. Union*, **25**(4): 533–537, 1944.

Heppner, J. P., A study of relationships between the aurora borealis and the geomagnetic disturbances caused by electric currents in the ionosphere, Ph.D. dissertation, California Institute of Technology, 125 pp., 1954.

———, Ness, N. F., Scearce, D. S., and Skillman, T. L., Explorer 10 magnetic field measurements, *J. Geophys. Res.*, **68,** 1–46, 1963.

TWENTY-FOUR-HOUR OBSERVATIONS OF AURORA AT THE SOUTHERN AURORAL ZONE

T. N. DAVIS* AND R. N. DEWITT

Geophysical Institute, University of Alaska

The west-to-east nighttime reversal in longitudinal auroral motions is observed at Byrd, Antarctica, as at other auroral zone stations. Because of the high geographic latitude of the station, it is possible to see the aurora throughout the day, and a second reversal in longitudinal motion is observed. This reversal, from eastward to westward motion, tends to occur approximately 10 hours later than the nighttime reversal. The morphology of the 'daytime' aurora is described here; it differs in several ways from that of the aurora commonly observed in the hours near local midnight at this and other auroral zone stations.

INTRODUCTION

Byrd, Antarctica (geographic coordinates 79°59′S, 120°01′W), is located at or just inside the southern auroral zone [see Hultqvist (1961) and Bond and Jacka (1962) for determinations of the auroral zone]. Of the established observing stations in the northern and southern hemispheres, Byrd is unique in that it lies very close to the auroral zone peak and is also at high geographic latitude. Therefore, by using this station during the southern winter solstice, it is possible to carry out observations of the visual aurora at the auroral zone throughout an entire day.

All-sky camera film taken at Byrd during the periods May–June 1960 and May–June 1961 has been examined, the emphasis being on determining the direction of auroral motions. The analysis was performed by using cinematic projection of the films; such a procedure tends to de-emphasize the short-lived and small-scale motions present and to emphasize the over-all motions of auroral forms and of irregularities within the forms. Usually it is found that at any one time there exists a single direction of motion common to all auroral forms in the field of view of the camera.

*Now National Academy of Sciences—NASA resident research associate at Goddard Space Flight Center, Greenbelt, Maryland.

RESULTS

The result of a determination of the percentage hourly occurrence of overhead auroras during one of the periods studied (May–June 1960) is shown in figure 1. The curve in figure 1 represents

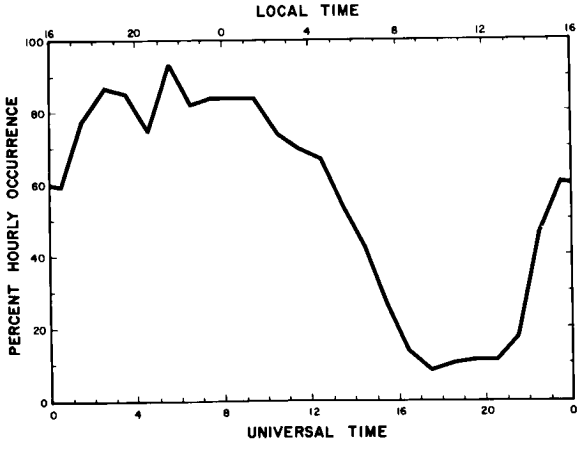


FIGURE 1.—Diurnal variation of hourly occurrence of visual aurora within 150 km of Byrd Station for all clear, dark hours in May and June 1960.

the occurrence of aurora within 150 km of the station. The auroral occurrence to the north and south of the Byrd zenith has been examined, and it was found that the occurrence was nearly sym-

metrical about the zenith. This result indicates that Byrd lies very close to the peak of the auroral zone; other characteristics of the morphology suggest that Byrd lies near and on the inner side of the auroral zone peak.

The results of the study of the auroral motions are summarized in figures 2 and 3. Figure 2 indicates the temporal distribution of both longitudinal and meridional motions in a format suggested by a similar diagram published by Evans (1960) to describe auroral motions at Halley Bay. Each data point indicates the predominant direction of the auroral motion during a fifteen-minute interval. The diagram shown in figure 3 pertains to longitudinal motions only. The pertinent points illustrated in the diagrams are mentioned in the following description of the different phases of the auroral display as observed at Byrd.

It is logical to begin the description with the phase of the display commencing near 2000 U.T.

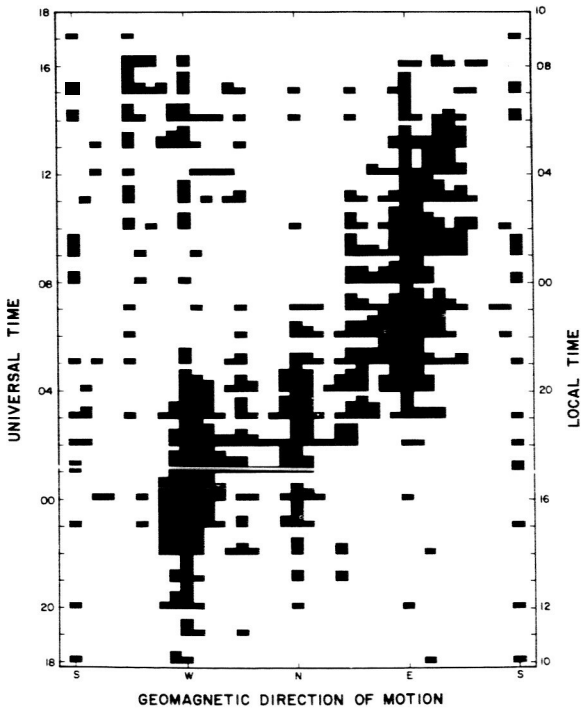


FIGURE 2.—Geomagnetic directions of horizontal auroral motions at Byrd Station. The direction shown is the direction toward which the aurora is moving. The diagram is compiled from all sky data from May–June 1960 and May–June 1961.

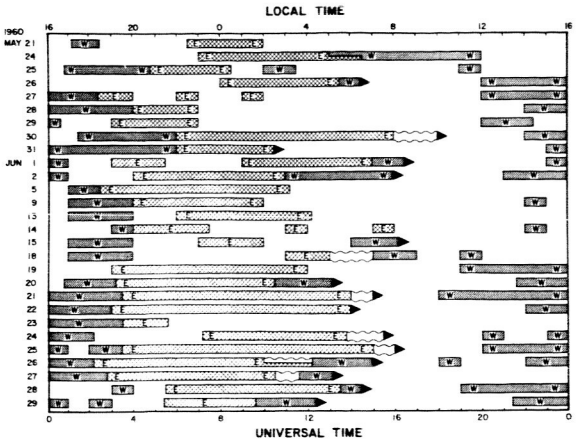


FIGURE 3.—Longitudinal auroral motions at Byrd. W and E represent the direction toward which the aurora moves. Parts of the diagram enclosed by wavy lines indicate period of no motion (see text), and dark triangles represent relatively sudden disappearances of the aurora.

This is a time of minimum auroral occurrence (figure 1).

2000–0200 U.T.

The auroral forms occurring during this interval are of the type normally observed at the auroral zone during evening hours. Homogeneous and rayed arcs and bands appear overhead or drift into view across the southern horizon. Pronounced westward drift occurs, both of irregularities within the forms and of the entire auroral forms. See figures 2, 3, and 4a.

0200–0700 U.T.

At the beginning of this interval westward motions continue and equatorward motions of arcs become more pronounced. The auroral forms usually become more numerous, and bandlike forms and looping structures often appear (figure 4b). A reversal to eastward longitudinal motion then occurs. The range in time of reversal of 27 observed events is 0105–0600, the median time being 0330. Near this time the arcs become greatly contorted, and meridionally aligned auroral forms frequently occur. This part of the display at Byrd corresponds to the breakup, which is accompanied by a change in type of auroral

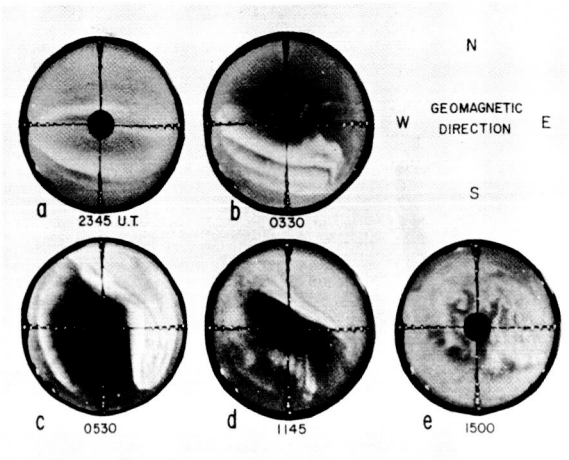


FIGURE 4.—All-sky photographs showing auroral forms typical of the periods near the times at which the photographs are taken. Photographs are taken at Byrd during June 1960: (a) homogeneous arcs typical of the forms during the evening phase; (b) bands of the type observed in the 0200–0700 U.T. interval; (c) meridionally aligned forms; (d) an example of the sequence from south to north of rayed forms, homogeneous arcs, diffuse aurora; and (d) interconnected diffuse forms during a period of no longitudinal motion.

form. The change in type is especially pronounced at somewhat lower latitude; at Byrd the change in type usually is not abrupt, but there is a tendency for the forms to be more diffuse after the reversal in longitudinal motion. After the reversal to eastward motion, the auroral forms tend to become less numerous.

The phases of the display discussed to this point are observable at any station having the same relationship to an auroral zone as Byrd. However, because of the occurrence of twilight the phase of the display described below is not observable at other established observing stations.

0700–2000 U.T.

The auroral forms continue to move primarily eastward at the beginning of this interval. In most instances a poleward component is added, so that the motion is directed toward east-southeast. As the forms drift over the station a definite sequence of form types is observed. Rayed arcs appear at the south edge of the display; diffuse homogeneous arcs lie to the north of the rayed arcs; north of these, diffuse surfaces occur (figure 4d). By 1000, the diffuse surfaces usually lie directly over Byrd and may completely cover the visible sky.

In those instances where observing conditions allowed full observation of the display during this interval, one of the following sequences was observed: (1) the eastward moving aurora dimmed and disappeared; or (2) the eastward motion ceased and the auroral forms remained overhead with no further motion for several tens of minutes (figures 3 and 4e); then the stabilized forms dimmed and disappeared in place; or westward motion developed before disappearance of the forms; or (3) a reversal from eastward to westward motion occurred without an intermediate phase of stabilized forms.

The range in time of 14 reversals from eastward to westward longitudinal motion (either with or without a period of motionless aurora) was 0930–1600, the median time being 1330. One striking feature of the morphology during this phase of the display is the disappearance of the aurora in the hours between 1200 and 1800; these disappearances are indicated on figure 3 by black triangles. Following the disappearance of the aurora, the sky over Byrd remained noticeably dark until the commencement of the phase of the display described in the paragraph above headed '2000–0200 U.T.'

SUMMARY AND DISCUSSION

Real time observation of auroral motions shows much small-scale complexity, yet the motions are observed to follow a rather simple pattern when viewed with a compressed time scale. The observations of auroral motions over Byrd show that this pattern extends over the auroral zone to all local times, permitting the auroral zone to be roughly divided into two segments, one on the sunset side of the earth where westward motion predominates and the other on the sunrise side of the earth where eastward motion predominates. On roughly the sunward part of the auroral zone the occurrence of aurora is a minimum.

In using the local geomagnetic time coordinate to compare the Byrd observations with those obtained at the northern auroral zone, we have found some disagreement which indicates that geomagnetic time is not a satisfactory coordinate

with which to discuss the southern observations. Nakamura (1963) has reached a similar conclusion from study of auroral observations taken at Syowa Base, which is also near the southern auroral zone.

ACKNOWLEDGMENT

This research was supported by the Atmospheric Sciences Program, National Science Foundation, NSF grant G14782.

REFERENCES

Bond, F. R., and Jacka, F., Distribution of auroras in the southern hemisphere, *Australian J. Phys.*, **15,** 261–272, 1962.

Evans, S., Systematic movements of aurorae at Halley Bay, *Proc. Roy. Soc. London*, A**256,** 234–240, 1960.

Hultqvist, Bengt, Circular symmetry in the geomagnetic plane for auroral phenomena, *Planetary Space Sci.*, **8,** 142–150, 1961.

Nakamura, Junji, The life and progress of the aurora, translated from *Kagaku*, **32,** 354–361, 1962, by E. R. Hope, *Defence Research Board*, T**78J,** 1963.

LOW-ENERGY TRAPPED PROTONS

LEO R. DAVIS

AND

JAMES M. WILLIAMSON

Goddard Space Flight Center

A scintillation detector designed to measure the directional intensity and spectrum of 100 kev to 4.5 Mev protons and the directional energy flux and spectrum of 10 to 100 kev electrons has been flown on the Explorer XII satellite (1961υ). Analysis of a portion of the data recorded during the last half of August 1961 has shown that (1) the radial distance to the outer boundary of the geomagnetically trapped radiation is variable, ranging from 8.5 to 11.0R_e; (2) electrons are trapped more or less uniformly from 2R_e to the outer boundary; and (3) protons are trapped throughout the same region.

The proton intensity peaks on the dipole field line having an equational radius (R_0 value) of about 3.5R_e, where the maximum intensity is 6×10^7 protons/cm^2-sec-ster. The proton spectra may be approximated by $\exp(-E/E_0)$ with, for example, E_0 values of 400, 120, and 64 kev at R_0 values of 2.8, 5.0, and 6.1R_e, respectively.

During the magnetic storm which began on September 30, 1961, the proton intensity measured at high latitudes was enhanced by a factor of 3. The possibility that the main phase decrease of the storm was produced by the protons is discussed.

INTRODUCTION

The directional intensity and spectrum of 100 kev to 4.5 Mev protons and the directional energy flux and spectrum of 10 to 100 kev electrons trapped in the geomagnetic field have recently been measured on the Explorer XII satellite (1961υ). Explorer XII was launched at0 321 U.T., August 16, 1961, into an elliptical orbit inclined 33 degrees to the equatorial plane, with a perigee altitude of 300 km, an apogee altitude of 77,300 km, and an orbital period of $26\frac{1}{2}$ hours. Initially, apogee was near local noon. The satellite instrumentation functioned properly up to December 6, 1961, when it abruptly ceased transmitting. Data were recorded nearly continuously throughout the life of the satellite, and when processed will be available for about 80 percent of that time. When the satellite ceased transmitting, apogee was near six hours local time.

The analysis, to date, has been in the nature of an initial survey to establish the proper operation and calibration of the instrumentation and to determine the gross features of the electron and proton radiation. The results have shown that (1) the trapped radiation exhibits a well marked outer boundary whose position is variable, ranging from 8.5 to 11.0R_e: (2) electrons are trapped more or less uniformly from 2R_e to the outer boundary; and (3) protons are trapped throughout the same region.

The present paper reports the initial results on the low energy trapped protons.

INSTRUMENTATION

The ion–electron detector flown on Explorer XII is similar to detectors flown previously on sounding rockets to measure auroral particles[1] and solar flare protons.[2] It employed a 4.4 mg/cm^2 thickness of powder phosphor, ZnS(Ag), settled on the face of a photomultiplier tube which was located behind a stepping absorber wheel. This is illustrated in figure 1, which also shows a block diagram of the electronics. Both the dc output and the pulse counting rate of the phototube are

telemetered. The ambient radiation can be admitted through either of the two apertures shown in figure 1, depending on the wheel position. When radiation entering through the lower aperture is allowed to pass through the wheel, it reaches the phosphor directly. Radiation entering through the upper aperture can only reach the phosphor after having scattered off the gold surface shown positioned in front of the apertures in figure 1.

When the detector is operated in the scatter mode, the dc current of the phototube is essentially a measure of the energy flux of electrons since few protons or other heavy particles will be scattered. In the direct geometry, the current is a measure of the total energy flux of particles stopping in the phosphor. The measured sensitivities as a function of particle energy are shown in figure 2 for electrons and protons incident in the direct geometry and for electrons in the scatter geometry. Sensitivity is defined here as the phototube output current per unit power of the monoenergetic particle beam passing through the aperture.

The trigger level of the pulse channel is set to count ions losing 80 kev or more energy in the phosphor. The measured proton counting efficiency as a function of proton energy is shown in figure 2. From the figure it can be seen that the proton counting efficiency exceeds 50 percent between 100 kev and 4.5 Mev and falls off rapidly outside this energy range. The detector also counts heavier ions with high efficiency, i.e., alpha particles having energies between 120 kev and 130 Mev. Electrons are discriminated against

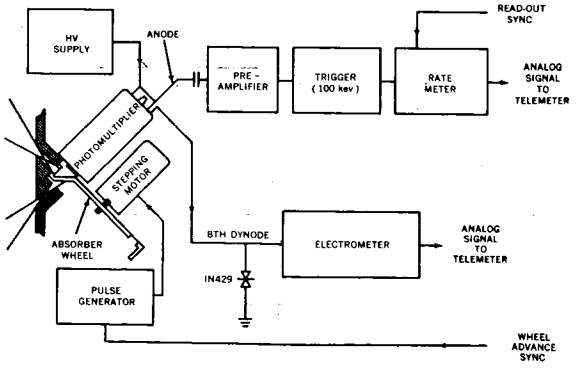


FIGURE 1.—Simplified drawing of ion-electron detector geometry and block diagram of electronics.

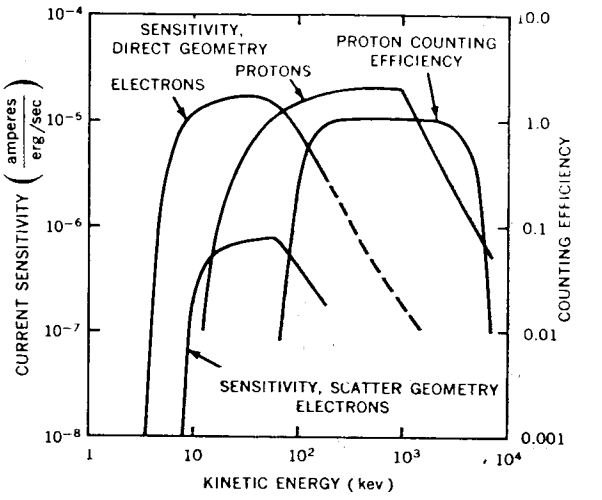


FIGURE 2.—Current sensitivities to electrons and protons and proton count efficiency of ion-electron detector with no absorber.

by the phosphor thinness (the maximum average energy loss of an electron which will just traverse the phosphor is 60 kev) and by the particular light-pulse-decay characteristic of the ZnS(Ag) phosphor which allows electron-produced pulses to be partially differentiated out in the pulse amplifier. As a result, single electrons will not be counted.

The low energy cutoff of the response curves shown in figure 2 results from the particles having to penetrate a 1000 A thick aluminum coating on the phosphor (and, in addition, for ions to lose 80 kev in the phosphor to count). The wheel is used to interpose additional absorber which moves the low energy cutoff to successively higher values. The ratio of outputs with the different absorbers gives a rough measure of the energy spectra and a quite sensitive measure of any change in the spectra of electrons or ions. A thick plug position is provided on the wheel to determine the background produced by any penetrating radiation reaching the phosphor through the sides of the detector. There are also three wheel positions with radioactive sources for inflight calibration.

This paper presents proton intensities and spectra derived from the counting rate and absorption data on the assumption that all of the particles counted were protons. In this regard it should be noted that a second count rate channel

was included to determine the ratio of protons to heavier ions. The trigger level of this channel was set greater than the energy loss of a proton which would just traverse the phosphor, i.e., at 900 kev compared to 800 kev. The results of a preliminary analysis show that a significant portion of the particles were heavier than protons. These results are dependent on the trigger level remaining stable, and an analysis which will test this stability is now being made. Should the results be confirmed, the present proton intensities will need to be corrected downward. The correction will be 30 percent or less, and the shape of spectra will be little affected.

The count-rate meter has a 0 to 5 volt analog output which is proportional to the logarithm of the number of counts stored since the last readout. The output was sampled three times per second. Telescope factors of 5.8×10^{-4} and 5.4×10^{-3} cm²-ster were employed and proton intensities up to 10^{7}/cm²-sec-ster could be measured.

The electrometer has a 0 to 5 volt analog output proportional to the logarithm of the input current for currents of 10^{-10} to 10^{-4} amp. Thus, electron and proton energy fluxes from about 10^{-2} to 10^{+4} erg/cm²-sec-ster could be measured.

The detector was positioned to look out at 45 degrees to the satellite spin axis. Full opening angles of 15 degrees and 22 degrees were used. The satellite spun about an axis which was fixed in inertial coordinates with about a 2-second spin period. Thus, it is possible to measure directly the angular dependence of the particle fluxes. Over the satellite orbit, the angle between the spin axis and geomagnetic field varied from 25 degrees to 90 degrees; therefore, the scanned range of particle pitch angles varied from as much as 0 to 90 degrees to as little as 20 to 70 degrees. The detector orientation in inertial coordinates and with respect to the local magnetic field was measured by a solar aspect indicator, and a three component magnetometer (the magnetic field measurements on Explorer XII were performed by Dr. Lawrence Cahill, of the University of New Hampshire, who kindly provided the magnetic orientation data).

RESULTS

Figure 3 shows counting rates and currents as a function of radial distance measured on

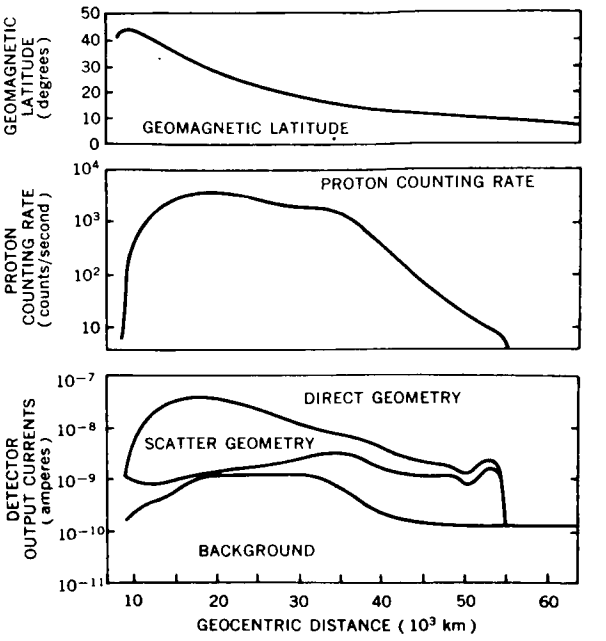


FIGURE 3.—Ion-electron detector output currents and count rate as a function of geocentric distance on Explorer XII pass 7b August 24, 1961. The geomagnetic latitude of the satellite is also shown.

August 24, 1961, for the wheel positions whose response curves are shown in figure 2. Also shown is the background current measured with the thick plug covering the apertures, and the geomagnetic latitude of the satellite. The counting rate curve shows that protons were present from

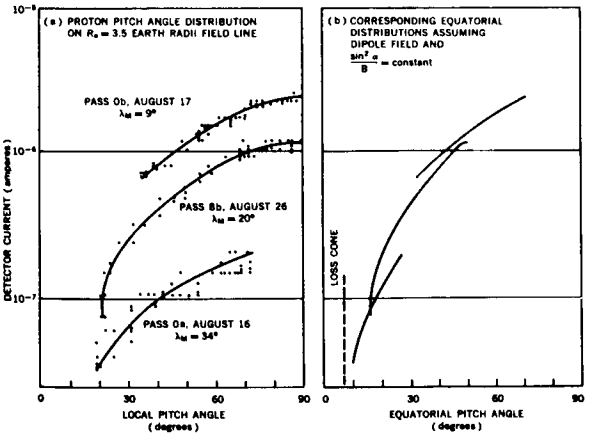


FIGURE 4.—(a) Measured local pitch angle distributions of low energy protons on three crossings of the 3.5 R_e field line; (b) The corresponding equatorial pitch angle distributions.

54,000 km radial distance down to 10,000 km. The peak proton intensity measured on this pass is 7×10^6/cm²-sec-ster at three R_e. The difference between the current measured in the scatter geometry and the background current is a measure of the electron energy flux. As may be seen, the electron flux rises to a maximum at 35,000 km and slowly decreases with increasing distance out to 54,000 km, where an abrupt discontinuity is observed and the electron flux drops below a detectable value. The difference between the current measured in the direct geometry and scatter geometry could only be produced by protons and thus confirms the count rate data. The peak in the background current apparently is due to the penetrating radiation which forms the "outer belt" as defined by Geiger–Müller counter measurements.[3]

The pitch angle distributions of the protons on the dipole field line having an equatorial radius (R_0 value) of $3.5R_e$ are shown in figure 4. In figure 4a are plotted the eight dynode currents measured at various local pitch angles on three crossings of the field line, each at a different geomagnetic latitude. The smooth lines drawn through the data points have been transformed from local pitch angle to equatorial pitch angle, assuming a dipole field and a constant $(\sin^2\alpha)/B$, and are shown in figure 4b. The three segments of pitch angle distribution so obtained are in reasonable agreement and together trace out the equatorial pitch angle distribution from 10 to 72 degrees. The resulting distribution is well approximated by a $\sin^3\alpha_0$ curve, where α_0 is the equatorial pitch angle.

Proton spectra measured on three field lines on August 26 are shown in figure 5. The data are plotted as integral spectra; however, it should be noted that the detector does not measure protons having energies greater than 4.5 Mev and we are thus assuming there is no significant number of the higher energy protons. The resulting steep spectra appear to justify this assumption. Spectra at a number of positions and times have been analyzed and, in general, the trapped proton spectra are better approximated by an exponential law than a power law. As may be seen, the spectra steepen as we move from lower to higher field lines. On field lines having R_0 values less than about 3 or $4R_e$, the spectra show a flattening

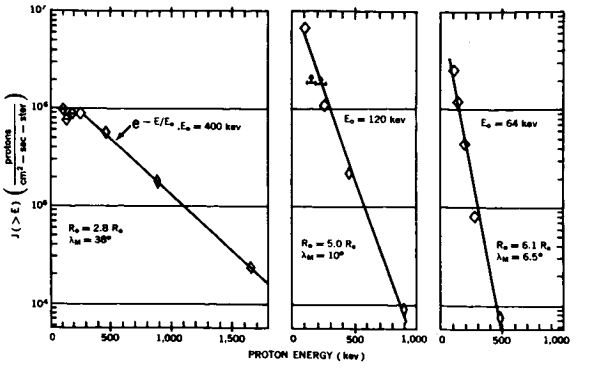


FIGURE 5.—Integral proton spectra measured on three field lines by Explorer XII (pass 8b), on August 26, 1961.

below a few hundred kev energy which means that protons below this energy are absent. This cutoff energy is about the energy below which the charge exchange process becomes dominant over the coulomb scattering process in determining the proton lifetimes. For example, the spectrum shown in figure 5 for $2.8R_e$ flattens between 200 and 300 kev. The lifetimes calculated by Liemohm[4] for this field line and latitude show that a 200 kev proton has a lifetime of about 5 days and the lifetime decreases rapidly at lower energies, whereas a 300 kev proton has a 30-day lifetime and the lifetimes of higher energy protons increase relatively slowly with increasing energy. By comparison in figure 5 it will be noted that the proton spectrum obtained on the $5R_e$ field line extends down to 100 kev. The lifetime of a 100 kev proton on this field line, limited by charge exchange, is about 50 days.

Figure 6 shows the preliminary results of mapping the spatial distribution of the low energy protons. In figure 6b are plotted the intensities of mirroring protons as a function of R_0 for five passes in late August. In figure 6a the trajectories for these five passes are plotted in the geomagnetic latitude–R_0 plane. The points on each trajectory where the proton intensity reached 10^7 protons/cm²-sec-ster are indicated by circles. The dashed line drawn through these points is, therefore, the corresponding proton intensity contour. The contour for an intensity value of 10^6 is similarly shown. As may be seen the resulting contours show that the low energy proton intensity is a maximum at about $3.5R_e$.

The pulse channel of the detector is in satura-

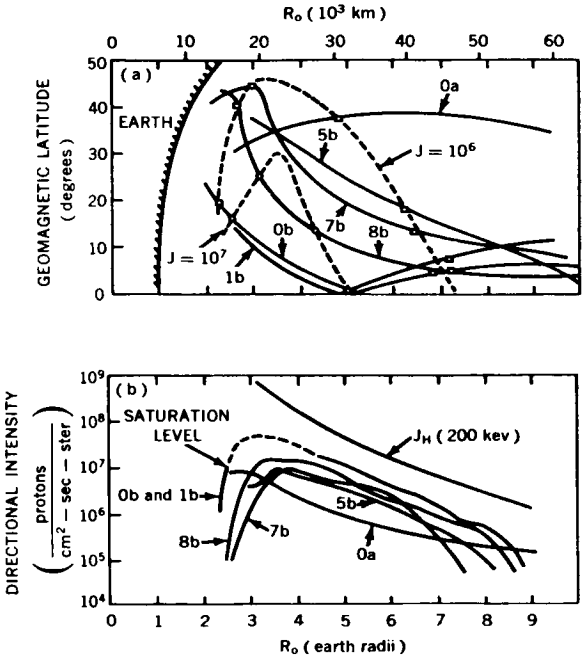


FIGURE 6.—(a) Six Explorer XII trajectories mapped in the geomagnetic latitude—R_0 plane, and proton intensity contours for $J=10^6$ and 10^7 protons/cm²-sec-ster; (b) Directional intensities of locally mirroring protons (120 kev$\leq E\leq$4.5 Mev) as functions of R_0 measured on six passes through the proton belt.

tion at intensities above 10^7/cm²-sec-ster. This effect can be seen on pass 8b in figure 6b. The dc channel at this time is still some three decades below saturation and thus may be used to measure the proton intensity at the higher values. This has been done for passes 0b and 1b in figure 6b. As may be seen, the maximum proton intensity is about 6×10^7/cm²-sec-ster measured some 7 degrees off the geomagnetic equator. This measurement should be accurate to $\pm$50 percent. A more accurate value will be obtained when the spectrum in this region is derived. Using the measured pitch angle distribution and assuming an average proton energy of 400 kev, we calculate the proton density to be one-half proton per cubic centimeter. The corresponding proton kinetic energy density of 2×10^5 ev/cm³ is about one-tenth the energy density of the geomagnetic field at this point.

The proton energy density in the region from 4 to $8R_e$ shows a remarkable tendency to track the field energy density. This is illustrated in figure 6b where the curve labeled "J_H(200 kev)" shows the directional intensity of 200 kev protons which, if isotropic, would have a kinetic energy density equal to the dipole field energy density. As may be seen, the proton intensity measured at low latitudes is about one-third of J_H (200 kev). When the proton pitch angle distribution is taken into consideration this means that the ratio of field energy density to proton energy density in the equatorial plane is approximately 6 in the region 4 to $8R_e$.

Nineteen more or less complete passes through the proton belt are presently available for studies of temporal variations. The data obtained on 15 of these passes, scattered through the period August 16 to September 30, 1961, show no temporal change in the proton intensities measured on field lines which have $R_0<7R_e$; i.e., the data are constant to within $\pm$30 percent.

At 2108 U.T. on September 30, a sudden commencement (SC) occurred which was followed by a magnetic storm lasting several days. At the time of the SC, Explorer XII was nearing apogee. The next pass through the proton belt occurred some 17 hours later. On this pass and three succeeding passes on October 1 and 2, the proton intensities measured in the region from 3 to $4.5R_e$ were greater by a factor of 3 than the pre-storm values. However, the satellite crossings of this region were at geomagnetic latitudes greater than 30 degrees so that only protons having equatorial pitch angles of 34 degrees and less were sampled. From the pre-storm pitch angle distribution, it is known that only 4 percent of the protons trapped in these field lines have equatorial pitch angles of 34 degrees or less. Thus, with these data we cannot determine whether the total proton population was increased or whether there was simply a redistribution in pitch angles of the pre-storm population.

The total kinetic energy of the trapped low energy protons exceeds that of any other known population of trapped particles, and thus their disturbance of the geomagnetic field is greatest. It is of obvious interest to determine the magnitude of the disturbances. Akasofu, Cain, and Chapman,[5] using a computer code developed for a previous theoretical study, have calculated the disturbance for a model proton belt which closely

approximates our pre-storm belt. The results predict a decrease in the surface equatorial field of 40γ. They have further calculated (private communication) the additional disturbance which would result if the storm time increase existed at all values of pitch angle. The results predict an 80γ additional decrease which is to be compared with D_{st} values of 60 to 30γ on October 1 and 2. Thus, we may hypothesize that the main phase decrease was produced by these low energy protons. A second large magnetic storm and several small storms occurred during the life of Explorer XII. When these data become available, further tests of this hypothesis will be made.

There are no previous measurements of trapped protons below 1 Mev in the region of the outer belt. However, the rocket measurements of Bame, Conner, et al.[6] have clearly shown that protons having energies down to 1 Mev are trapped on field lines having R_0 values around 2.5 to 3R_e. A preliminary comparison with their data, using, however, only dipole coordinates, shows agreement in both intensity and spectral slope for proton energies of around 1 Mev. Our results may be related to the results of Naugle and Kniffen[7] and of Freeman[8]; however, to date no comparison has been made.

REFERENCES

1. Davis, L. R., Berg, O. E., and Meredith, L. H., "Direct Measurements of Particle Fluxes In and Near Auroras," in: *Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January* 1960, ed. by H. K. Bijl, Amsterdam: North-Holland Publ. Co., pp. 721–735, 1960.
2. Ogilvie, K. W., Bryant, D. M., and Davis, L. R. "Rocket Observations of Solar Protons during the November 1960 Events, 1," *J. Geophys. Res.*, **67**(3): 929–937, March 1962.
3. O'Brien, B. J., Van Allen, J. A., et al., "Absolute Electron Intensities in the Heart of the Earth's Outer Radiation Zone," *J. Geophys. Res.*, **67**(1): 397–403, January 1962.
4. Liemohn, H., "The Lifetime of Radiation Belt Protons with Energies between 1 Kev and 1 Mev," *J. Geophys. Res.*, **66**(10): 3593–3595, October 1961.
5. Akasofu, S. R., Cain, J. C., and Chapman, S., "The Magnetic Field of the Quiet-Time Proton Belt," *J. Geophys. Res.*, **67**(7): 2645–2647, July 1962.
6. Bame, S. J., and Conner, J. P., et al., "Protons in the Outer Van Allen Belt," *J. Geophys. Res.*, **67**(4): 1628, April 1962 (Abstract).
7. Naugle, J. E., and Kniffen, D. A., "Flux and Energy Spectra of the Protons in the Inner Van Allen Belt," *Phys. Rev. Letters*, **7**(1): 3–6, July 1, 1961; also NASA Technical Note D–412, August 1961.
8. Freeman, J. W., "Detection of an Intense Flux of Low-Energy Protons or Ions Trapped in the Inner Radiation Zone," *J. Geophys. Res.*, **67**(3): 921–928, March 1962.

VERY HIGH-ENERGY COSMIC RAYS

C. E. FICHTEL

Goddard Space Flight Center

Recently, Linsley and Scarsi[1] have shown that present evidence indicates that the primary particles of total energy $\gtrsim 10^{17}$ eV are nearly all protons. They further state that the heavy nuclei are less abundant than at lower energies, and may be completely absent. In a companion paper, Linsley[2] suggests that one possible explanation of this feature is that the cosmic rays of lower energy are of galactic origin, those above about 5×10^{16} eV are of metagalactic origin, and at about 10^{16} eV/nucleus the two sources make an approximately equal contribution. The intent of this Letter is to show that there is another possible interpretation, which is based on the proposition that spiral-arm segments in the galaxy may be an important controlling feature. This hypothesis can also account for the other principal known features of low- and high-energy cosmic rays.

Peters[3] has already discussed some of the important features of the cosmic radiation which can be explained by considering the effect of the spiral arms. To summarize the basic ideas of this model, omitting the refinements, the particles, after being accelerated in the supernovae, travel along a spiral arm until the particles reach the end of the arm, or the arm segment, and then enter into the galactic halo. Peters shows that the cosmic rays formed in supernovae could pass through only a few g/cm^2 of interstellar material before leaving the arm segment, in agreement with the amount of matter traversed as deduced from the relative abundance of the light nuclei ($3\leq$nuclear charge≤5) in cosmic rays. He further notes that it may be reasonable to assume that as the particles reach the end of the arm segment they diffuse into the halo, where their average lifetime could be quite large, perhaps approaching half the age of the galaxy. He then points out that since they pass through a relatively large amount of material there will be a decrease in the number of complex nuclei due to fragmentation.

A look will now be taken at the further implications of this model. Because of the finite size of the spiral-arm segment and the strength of the field therein, there must be an upper limit to the rigidity of a particle which can be held easily in a spiral-arm segment. If a radius of curvature which is one-hundredth that of the estimated diameter of a spiral-arm segment is chosen as an upper limit for particles which should be held fairly easily in the galactic arm, the corresponding rigidity is given by

$$R_{\rm limit}=300aH=(300)(9\times10^{18})(6\times10^{-6}), \quad (1)$$

$$R_{\rm limit}=1.5\times10^{16}\ \text{eV}; \quad (2)$$

where R is the rigidity in eV, a is the radius of curvature in cm, and H is the field strength in gauss.

Particles with rigidities above about this value will escape from the spiral arm relatively easily and subsequently pass through them relatively easily. If the average amount of material through which they pass is calculated on the basis of the assumption that the amount of time they spend in the galactic disk and the halo is proportional to their relative volume, the following expression is obtained:

$$\bar{\rho}=\left[\frac{V(\text{disk})}{V(\text{disk})+V(\text{halo})}\right]\rho(\text{disk}) + \left[\frac{V(\text{halo})}{V(\text{disk})+V(\text{halo})}\right]\rho(\text{halo}). \quad (3)$$

ρ(disk) and ρ(halo) are assumed to be 10^{-24} and 10^{-26} g/cm^3, respectively.[4,5] V(halo) and V(disk) are calculated as follows:

$$V(\text{halo})=(4/3)\pi(4.5\times10^{22})^3=4\times10^{68}\ \text{cm}^3, \quad (4)$$

$$V(\text{disk})=\pi(4.5\times10^{22})^2(9\times10^{20})=6\times10^{66}\ \text{cm}^3. \quad (5)$$

From (3), (4), and (5),

$$\bar{\rho}=2.5\times10^{-26}\ \text{g/cm}^3. \quad (6)$$

The total amount of material through which these ultrahigh-rigidity particles have passed in a period of time equal to half the estimated age of the galaxy[5] is then given by

$$(2.5\times10^{-26}\ \text{g/cm}^3)\,(3\times10^{10}\ \text{cm/sec})\,(2\times10^{17}\ \text{sec}) = 150\ \text{g/cm}^2. \quad (7)$$

Since this is very large compared to the mean free path of absorption for heavy nuclei, about 8 g/cm^2, and even helium nuclei, about 16 g/cm^2, high-rigidity particles should consist almost exclusively of protons. Since in terms of the cosmic-ray power-law spectrum in total energy protons, helium nuclei, and heavy nuclei are known to have nearly the same relative abundance above a given total energy at lower energies,[4] $\leq10^{13}$ eV, the transition at about 10^{16} eV rigidity corresponds to a very marked transition in composition at about 10^{16} to 10^{17} eV total energy if the source composition distributions continue to hold at energies above 10^{13} eV as would be expected.

At the same time, local scattering in the spiral arm probably becomes less and less important as the rigidity increases, so that as the escape rigidity from the spiral-arm segment is approached the motion of the particle would be primarily controlled by the general field parallel to the spiral arm. An anisotropy relative to the angle of the particles with respect to the axis of the spiral arm for primary heavy nuclei in the energy range from 10^{16} to 10^{17} eV/nucleus, such as the one suggested by the data of Hasegawa et al.,[6] might then be expected.

Because of the different history of the low- and very high-energy particles, a break in the energy spectrum of cosmic rays might also be expected at about 10^{16} eV. If the differential energy spectrum at the source is continuous, then the break will consist principally of a change in intensity and not spectral shape at this point with the change in intensity being determined primarily by the volume and lifetime of particles in the whole galaxy relative to those in the spiral arms. The volume of all the spiral-arm segments will be estimated as $10\pi r^2 l=10\pi(4.5\times10^{20})^2\times(2\times10^{22})=1.3\times10^{65}$ cm^3, and the lifetime in the arm segment as the time necessary to travel 4 g/cm^2 in a medium of density 10^{-24} g/cm^3, namely, 4×10^6 years. Combining these numbers with the corresponding values for the halo given earlier, an intensity decrease in the differential spectrum of a factor of about two is expected, although this number is uncertain by as much as a factor of ten. A small change in the differential energy spectrum at this point in energy is consistent with the experimental data.[7]

It seems less likely that there should be such a smooth transition in the energy spectrum if the change in cosmic-ray composition at about 10^{16} to 10^{17} eV is attributed to a change from galactic to metagalactic origin. There must, of course, be some rigidity above which particles cannot be held easily in the galaxy, but it could reasonably be expected to be about two orders of magnitude larger than that necessary to hold a particle in an arm segment. This higher rigidity, which would then be of the order of 10^{18} to 10^{19} eV rigidity, will correspond to the transition from galactic to metagalactic particles. Another change in the differential energy spectrum might then be expected at this point.

The model considered here leads to a change in composition and a small change in the energy spectrum at about the same energy, estimated to be about 10^{16} to 10^{17} eV, with the change in composition occurring about a factor of two to ten higher in total energy. In addition, a possible heavy anisotropy relative to the spiral-arm segment at these energies can be explained. Finally, another and probably more significant change in the differential energy spectrum is expected at about 10^{19} eV total energy.

REFERENCES

1. Linsley, John and Scarsi, Livio, *Phys. Rev. Letters*, **9**, 123 (1962).
2. ———, *Phys. Rev. Letters*, **9**, 126 (1962).
3. Peters, B., *Nuovo Cimento, Suppl.*, **14**, 436 (1959).
4. Ginzburg, V. L., and Syrovatsky, S. I., *Progr. Theoret. Phys. (Kyoto), Suppl. No.* **20**, 1 (1962).
5. Shapiro, M. M., *Science*, **135**, 175 (1962).
6. Hasegawa, H., Matano, T., Miura, I., Oda, M., Tanahashi, G., Tanaka, Y., Higashi, S., Kitamura, T., Mishima, Y., Miyamoto, S., Shibata, K., and Watase, Y., *Phys. Rev. Letters*, **8**, 284 (1962).
7. Waddington, C. J., *Progr. Nucl. Phys.*, **8**, 13 (1960).

COSMIC-RAY HYDROGEN AND HELIUM NUCLEI DURING A SOLAR QUIET TIME IN JULY 1961

C. E. FICHTEL AND D. E. GUSS

Goddard Space Flight Center

AND

G. R. STEVENSON* AND C. J. WADDINGTON†

University of Bristol

The energy spectra of low-energy primary cosmic-ray hydrogen and helium nuclei have been determined from nuclear emulsions flown on a high-altitude balloon launched from Fort Churchill on July 8 1961. The flight was preceded by a period of three months of low solar and geophysical activity. The results indicated that the differential spectrum of hydrogen nuclei in this low-energy interval decreased toward low energies and did not exhibit the sharp rise found with counters in 1960 and 1961 by other experimenters. The low-energy differential spectrum and integral flux value for helium nuclei were found to be intermediate between those typical of solar minimum and solar maximum and are consistent with a modulation mechanism which yields the same rigidity spectrum for a given intensity both before and after solar maximum. Various modulation models of galactic cosmic rays are examined, and it is concluded that none of them seem to be entirely adequate.

INTRODUCTION

The intensity of the cosmic radiation detected at the earth is subject to a modulation that shows a strong inverse correlation to the eleven-year cycle of solar activity.[1,2] The changes in intensity over the solar cycle are most pronounced for particles of low magnetic rigidity, and it is data from these particles which should yield the most useful information concerning the mechanism of the modulation. In particular, a comparison of the energy or rigidity spectra of particles which have different rigidities for the same velocity, such as hydrogen and helium nuclei, must provide rigid restrictions on any proposed modulating mechanism.

During the period from 1955 to 1960 McDonald and Webber[3–6] made a series of measurements of the proton rigidity spectrum using Čerenkov-scintillation counter telescopes flown on balloons. These spectra were measured between the rigidity of the geomagnetic threshold of the individual balloon flights and about 1.5 BV. In addition the integral flux above 1.5 BV was determined. Although the work of Quenby and Wenk[7] predicts thresholds of 0.7 and 1.3 BV for the different individual balloon flights at high latitudes, the true threshold may be lower and may not be sharp. For these reasons, the lower end of the observed spectrum is difficult to interpret in terms of the spectrum in free space.

During a longer period which overlaps that of these measurements, the low-energy helium nuclei have been examined by many observers using nuclear emulsions.[8–16] In some instances the air cutoff rigidity, typically 0.7 BV for He nuclei, was higher than the geomagnetic threshold, and

*Present address: Radiological Protection Service, Clifton Avenue, Belmont, Sutton, Surrey, England.

†National Academy of Sciences, NASA senior postdoctoral resident research associate while at Goddard Space Flight Center. Present address: School of Physics, University of Minnesota, Minneapolis, Minnesota.

in almost every case, data were available to about 2.5 BV and sometimes to 4.5 BV. Additional information is provided by the results of Winckler and Anderson[17] and Neher[18] at northern latitudes in 1954 and 1955 which indicate a higher intensity of low rigidity particles at solar minimum than that which would be deduced by extrapolating the results of McDonald and Webber to lower rigidity. There is then the possibility of an additional increase in the proton flux below about 0.5 BV rigidity during solar minimum.

The following conclusions can be drawn from these results. (a) The intensities of the hydrogen and helium nuclei above about 0.7 BV decrease appreciably from solar minimum to solar maximum. (b) If the degree of transparency in the penumbral region is assumed to depend only on rigidity, these nuclei appear to have similar differential rigidity spectra with a hydrogen to helium nuclei ratio of about 7.0 at least for rigidities above about 1 BV. (c) The helium nuclei, and, if (b) is correct, the hydrogen nuclei also, have a differential rigidity spectrum with a peak intensity at about 1.8 BV at solar minimum and about 2.2 BV at solar maximum.

Because of the present uncertainty of the shapes of the low end of the rigidity spectra at solar minimum, it is important as solar minimum is again approached to study these spectra with instruments flown at high altitude and latitude that are capable of directly detecting and identifying primary particles. Vogt[19] and Meyer and Vogt[20] have made such measurements for protons in 1960 and 1961. At times which they associate with quiet day conditions, they found differential proton spectra that increased toward lower rigidities in a range from 0.38 to 0.63 BV. Furthermore, they found that in 1961 the intensity in this rigidity range had decreased from that observed in 1960. If these low rigidity protons are of galactic origin, one might expect an increase in the intensity as solar minimum is approached and the modulation weakens. The observed spectral shape would also impose an extremely severe constraint on the modulation mechanism. Meyer and Vogt, therefore, suggest that these low-energy protons are mostly of solar origin.

The data reported here are from a time when the sun was comparatively quiet and had been so for several months. It is of interest to compare these proton data with those of Vogt[19] and Meyer and Vogt[20] to see how the low-energy proton component varies with time. Also, in principle, a comparison of the rigidity spectra of the hydrogen and helium nuclei would show whether the apparently similar spectra observed by McDonald and Webber at higher rigidities continues down into this low rigidity region. The results of this latter comparison are ambiguous, however, because the presence of any solar injected particles will alter the abundance ratio. Finally, for the helium nuclei it is important to determine, by comparison with the previously published data how the spectrum varies, since this information will place restrictions on the modulation mechanism.

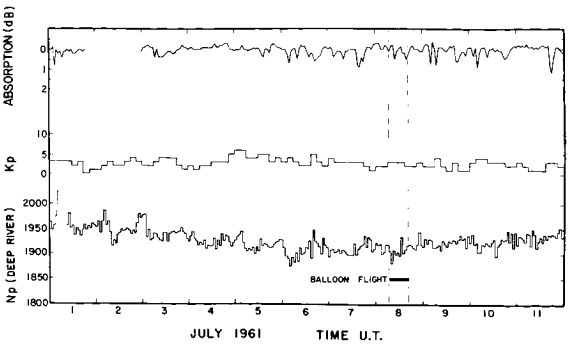


FIGURE 1.—Geophysical data during early July 1961. From top to bottom are shown absorption of 30-Mc/sec riometer, Fort Churchill; magnetic K_p index; and neutron monitor counting rate, Deep River.

EXPERIMENTAL DETAILS

Balloon Flight

The balloon flight on 8 July 1961, was made at a time of low solar activity preceded by a period of three months during which no significant Forbush or polar cap absorption events were observed. The last solar particle event detected by riometers preceding the flight was that on 21 November 1960. Figure 1 shows the cosmic noise absorption of the Churchill 30-Mc/sec riometer,[21] the counting rate of the Deep River neutron monitor,[22] and the geomagnetic 3-h range indices[23] K_p for the early part of July 1961. As can be seen, there were no signs of unusual activity prior to the flight nor

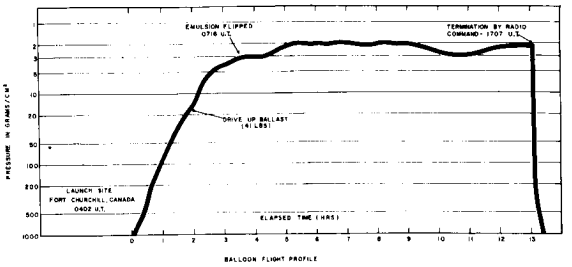


FIGURE 2.—Balloon flight profile.

after the flight until the solar particle event of 11 July. It would appear, then, that the particle intensities obtained from this balloon flight are truly indicative of undisturbed conditions appropriate to this period of the solar cycle.

The balloon flight path was essentially straight west from Fort Churchill and was at all times at a latitude where the air "cutoff" energy was considerably greater than the geomagnetic threshold for both protons and α particles. The balloon altitude profile is shown in Fig. 2. The pressure below 20 g/cm^2 was measured to within ± 0.1 g/cm^2 by photographing a Wallace and Tiernan (0–20 mm Hg) gauge.

The detector consisted of a large stack of 20-cm $\times$ 10-cm $\times$ 600-μ-thick Ilford emulsions of various sensitivities, of which 20 at one end were electron sensitive G5 emulsions. This stack was rotated through 180° at the ceiling altitude, and a supplementary emulsion package, the drop stack, intended to be used for the ascent correction, was released at the same time. This stack was unfortunately not recovered, and in its place drop stacks from a balloon flight at Sioux Falls, South Dakota, on 5 December 1961, and a balloon flight at Fort Churchill on 4 August 1962 were used for portions of the ascent correction.

Proton Analysis

The emulsions used in the proton analysis were the 20 G5 emulsions from the main flight stack and G5 emulsions from the Sioux Falls and Churchill drop stacks. Five different scans were made: (1) 1 cm from the top edge of the Churchill main stack, (2) 1 cm from the bottom edge of the main stack, (3) 1 cm from the bottom edge of the Sioux Falls drop stack, (4) 0.2 cm from the top edge of the main stack, and (5) 0.2 cm from the bottom edge of the Churchill drop stack. The top of the main stack faced the top of the atmosphere after rotation and faced the earth prior to rotation. The bottoms of the drop stacks faced the earth, and hence were oriented in the same way as the top of the main stack during balloon ascent. The scans were made to include tracks that lay within a pre-set solid angle. In no case did the zenith angle exceed 32°. Tracks with all grain densities were accepted in the first three scans while only those with grain densities ≥ 3 times the proton minimum were accepted in the last two.

Tracks from the first three scans were selected for subsequent analysis if they had an ionization greater than 1.8 times the proton minimum. Tracks due to mesons, tracks (except those which ended) which could not be followed out of the scan plate (and hence were produced before the stack was assembled), and tracks of multiply charged particles were rejected. The remaining tracks, which were due to singly charged particles, were followed in both directions from the scan line until they ended, interacted, or left the stack, and their direction of motion was determined, if possible. The energy at the scan line of each particle that did not end in the stack was determined by making blob-gap counts of the ionization,[24] and using a calibrated relation between these counts and the proton residual range. Each ionization estimate had a precision of about 5%.

The tracks from the scan 1 cm from the top of the main stack fall into three classes: (I) tracks of particles whose direction of motion was toward the bottom of the stack and which did not arise from an interaction in the emulsion between the scan line and the blackened edge of the emulsion, (II) tracks of particles whose direction of motion was toward the bottom of the stack but which originated in an interaction between the scan line and the blackened edge of the emulsion, and (III) tracks of particles whose direction of motion was toward the top of the stack.

In addition to true primary particle tracks which traversed the residual atmosphere above the balloon and entered the top of the emulsion stack while it was at the ceiling altitude, class I includes tracks of particles which entered the stack from underneath during the balloon ascent while

the stack was inverted, tracks which were formed on the ground after the stack was assembled and before and after it was flown, and tracks of particles which were produced in interactions between the blackened edge of the emulsion at the top of the stack and top of the atmosphere.

The corrections for that portion of the ascent up to a residual atmosphere of 4.5 g/cm^2 and for tracks formed in the stack before and after the flight were made directly from the scans in the Sioux Falls drop stack. The ascent between 4.5 g/cm^2 and 3.5 g/cm^2 was not covered by the drop stack, and the correction for this contribution was made from the scans at the bottom of the main stack. While the balloon was rising from 4.5 g/cm^2 to 3.5 g/cm^2, where the stack was rotated 180°, the top of the stack was looking toward the ground and particles which entered the stack at that time were produced in interactions below the stack. This, however, is the same component that was seen by the bottom of the stack for the remainder of the flight, after the stack rotation, when the balloon was at only a slightly higher altitude. Since the intensities for the bottom of the stack are very similar to those found from the top of the stack scan, and since the correction is small, one can approximate the correction sufficiently well by merely multiplying the bottom of the stack intensity corrected for ascent by a factor which is the fraction of time from 4.5 g/cm^2 to 3.5 g/cm^2 before stack rotation divided by the total flight time from 4.5 g/cm^2 to cutdown.

The remaining correction, that for particles from interactions in the atmosphere above the balloon and in the blackened edge of the emulsion, was made using the tracks from interactions observed in the emulsion above the scan line. It was assumed that the energy spectrum of secondary grey track[25] particles is the same for interactions with air nuclei as for interactions with emulsion nuclei. The internuclear cascade calculations of Metropolis et al.[26,27] and Bertini[28] show that this is a reasonable assumption in the range of secondary particle energies considered here. The energy lost by a particle in the amount of emulsion in which these interactions above the scan line are observed was almost identical to that lost in the amount of air above the balloon. Hence, the energy spectrum, but not the intensity, of these particles at the scan line is the same as the spectrum at the top of the stack produced by interactions in the residual atmosphere above the balloon. Once the intensity of this secondary component has been determined, as described below, the pseudospectrum of the particles arising from observed interactions in emulsion with energies corrected from the scan line through an amount of air equivalent to the amount of air above the balloon at ceiling and with the intensity normalized can be subtracted from the spectrum of particles of class I, corrected for ascent, and with energies corrected to the top of the atmosphere to yield the true primary spectrum.

The ratio of the intensity of particles from interactions in the atmosphere plus the obscured edge of the emulsion to that from observed interactions in the emulsion is given by the following equation:

$$S=\left(\frac{N_{ga}}{N_{ge}}\right)\left(\frac{n_{pa}+I_{pa}\,f_a n_{aa}}{n_{pe}+I_{pa}\,f_e n_{ae}}\right)+\left(\frac{x_{eu}}{x_{e0}}\right), \qquad (1)$$

where N_{ga}/N_{ge} is the ratio of the number of grey track secondaries per interaction in air nuclei to that in emulsion nuclei; n_{pa} and n_{pe}, the number of interactions per incident proton in air and in emulsion for this experiment; n_{aa} and n_{ae} the same, but for α particles; $I_{pa}=0.133$, the ratio of the primary alpha intensity to the primary proton intensity; f_a and f_e the ratios of grey track secondaries from proton interactions to that from alpha-particle interactions in air and in emulsion, and X_{eu} and X_{e0} are the depths of emulsion in which interactions were obscured by edge blackening and in which they were observed, respectively. The various values of n_{ij}, the number of interactions per incident particle are given by $n_{ij}=1-\exp(-x_j/\lambda_{ij})$, where x_j is the absorber thickness and λ_{ij} is the interaction mean free path. The values of λ_{ij} used here are $\lambda_{pa}=100$ g/cm^2, $\lambda_{aa}=45$ g/cm^2, $\lambda_{pe}=38$ cm, and $\lambda_{ae}=19.3$ cm.

The most uncertain parameters are the ratio N_{ga}/N_{ge}, f_a and f_e. There is very little experimental information available from which the ratio N_{ga}/N_{ge} can be determined.

Metropolis et al.[26,27] from a Monte Carlo calculation on internuclear cascades have obtained extensive data on the secondary cascade nucleons emitted in the bombardment of a number of

target elements ranging from aluminum to uranium with incident protons in the range of incident proton energies from 82 MeV to 1.8 BeV. Bertini[28] has recently completed a similar calculation including elements down to carbon and incident proton energies up to 400 MeV. The curves of average number of cascade protons per interaction versus atomic number of the target nucleus of these two calculations are similar in shape and the results of Bertini indicate that the average number of cascade protons per interaction of Metropolis et al. can be reasonably extrapolated to air nuclei.

The value of N_{ga}/N_{ge} can be found from these data by summing the average number of cascade protons per interaction as a function of incident proton energy for air nuclei and emulsion nuclei over the cosmic-ray proton spectrum. The result of this calculation was $N_{ga}/N_{ge}=0.85$, and, this is the value which was used in the calculation of S. Though this result is sensitive to the shape of the curve of average number of cascade nucleons versus atomic number, it is not sensitive to the absolute number of cascade protons per interaction, because it is a ratio.

A rough experimental estimate of N_{ga}/N_{ge} can be obtained from emulsion data by comparing the number of grey track secondaries for interactions characterized by $N_h \leq 7$ with the number of grey track secondaries from all interactions in emulsion[29] Lock et al.[30] studying interactions of 950-MeV incident protons in emulsion find that within limited statistics the number of grey plus shower particles is independent of N_h. Though this result was quoted for interactions on heavy nuclei only, the method of separation into interactions on heavy nuclei and on light nuclei was biased in such a way that almost no interactions with $N_h \leq 3$ were classified as interactions on light nuclei. However, Brown[31] studying star prong distributions in a nitrogen-filled cloud chamber at mountain altitude, found that most of these interactions had $N_h \leq 3$. Further, since the separation of Lock et al. excluded only 10% of the total number of interactions from being heavy nuclei, one could reasonably apply their result to all interactions in emulsion. Assuming then the interactions with $N_h \leq 7$ to represent air nuclei and that the average numbers of grey and shower particles are individually independent of N_h, the data of Lock et al. would indicate that $N_{ga}/N_{ge} \simeq 1$ for 950-MeV incident protons. This is completely consistent with the value deduced from the cascade calculations of Metropolis et al. The results of Beliakov et al.[32] and Zhdanov et al.[33] from interactions of 9-BeV protons in emulsion indicate a ratio of $N_{ga}/N_{ge} \simeq 0.5$. This result is consistent with an extrapolation of the data of Metropolis et al. to higher energy.

The values of f_a and f_e are not known, but are probably close to unity as evidenced by the fact that the average value of N_h for interactions in emulsion produced even by incident nuclei with $Z \geq 20$ is only 8.3 as compared with 8.0 for proton-induced interactions.[34] The values chosen here were $f_a=f_e=1$.

Tracks in the scan 0.2 cm from the top of the main stack were followed to a distance of 1 cm from the top of the stack and only tracks which ended were accepted. These tracks were those of protons with energies between 59 and 74 MeV at the top of the atmosphere. The same corrections were required as for the scans at greater depth in the emulsion and were made in the same way. The ascent correction was made using the scans 0.2 cm from the bottom edge of the 1962 Churchill drop stack emulsions and the scans at the bottom of the main Churchill flight stack, and the

TABLE I.—*Differential Hydrogen Nuclei Fluxes. (All intensities are in particles/m² sr sec MeV.)*

Energy interval at top of atmosphere (MeV)	59–74	80–160	160–250
Uncorrected intensity	1.40±0.36	0.77±0.12	0.81±0.11
Intensity corrected for ascent	0.98±0.39	0.63±0.15	0.69±0.14
Intensity at top of atmosphere (corrected for ascent and for secondary production in the atmosphere above the balloon)	0±0.5	0.34±0.17	0.50±0.16

secondary contributions from interactions above the stack were calculated from the tracks from observed interactions above the scan line found in scan 1.

The differential fluxes of hydrogen nuclei are shown in Table I. The uncorrected intensity and the intensity corrected for ascent alone are also given in order to show the magnitude of the corrections which are made to the raw data. The uncertainty in the ratio N_{ga}/N_e is not included in the listed error. No attempt was made to separate deuterons or tritons, and all of the energies calculated at the top of the atmosphere assume that the particles are protons.

He-Particle Data Reduction

In order to obtain the basic data for the helium nuclei, Ilford G5 nuclear emulsion plates from that part of the stack where the G5 plates were interleaved with G2 plates were scanned along a line parallel to the top edge of the stack for all tracks above a minimum grain density within a specified solid angle. In addition to a set of scans to determine the integral intensity and low-energy end of the energy spectrum, additional scans were performed specifically to obtain more information on the low-energy helium nuclei so that the modulation phenomenon could be studied with greater statistical accuracy. In both instances, the minimum grain density to be accepted in the initial scan was set about three standard deviations below the lowest value to be used in the analysis. The usual scanning efficiency checks[35,36] such as comparing the distribution of the numbers of tracks found as a function of the zenith angle, the dip angle, the depth in the emulsion, and grain density with the expected distributions and rescanning of about 30% of the total area by a second scanner were performed. In addition, the finally accepted solid angle was reduced from the original one set for scanning in order to avoid possible scanning loss near the limits even though none seemed to be present. With these precautions, the maximum correction for scanning loss was only 2%. All tracks coming from interactions above the scan line were rejected.

The helium nuclei were separated from singly and multiply charged nuclei by visual observation of the scattering and grain density or by measurement of these quantities if necessary. The energy of the helium nucleus which produced the track was then determined from the range of the particle in the stack if it ended, or from its ionization in a G2 emulsion if it did not. The ionization curve was determined in three different ways. For all tracks, a calibration was obtained by a comparison with the energy determined from multiple scattering measurements at high energies and range at low energies. In measuring the mean angle of deflection, the sagitta method[37] was employed with the noise elimination procedure of Menon et al.,[38] and the scattering constants calculated by Fichtel and Friedlander.[39] In addition, 30% of tracks with energies less than about 600 MeV/nucleon were analyzed by the method developed by Freier et al.[40] wherein it is assumed on the basis of theoretical considerations that the ionization is only a function of Z^2/β^n in the region of interest, namely, from 100 to 600 MeV/nucleon. A good fit for the set of data considered for protons and helium nuclei was obtained for $n=1.75$. The remaining 70% of the tracks were analyzed by assuming the Fowler-Perkins ionization curve[24] for protons multiplied by Z^2 and corrected for saturation, held for helium nuclei over the same energy region. Good agreement was obtained with the energies estimated from scattering data. In both cases it was noted that helium nuclei with appreciably higher energies, energies/nucleon $\gtrsim 1$ BeV, seemed to have a higher grain density than would have been predicted.

All helium nuclei whose grain density indicated that they might stop in the stack[41]; i.e., whose energy was less than about 200 MeV/nucleon, were followed until they stopped, interacted, or did leave the stack. In general, in addition to having a better estimate of the energy for particles with energies less than 200 MeV/nucleon, their direction was also known. This fact aids in the analysis to be described below.

In the whole of this calculation it has been assumed that the particles were He^4 nuclei. This assumption is based on the small percentage, ~20%, of He^3 observed in the experiments of Hildebrand et al.,[42] and Foster and Mulvey.[43] The experimental effect on the differential energy spectrum produced by as much as 50% of He^3

TABLE II.—*Integral Helium Particle Flux as a Function of Kinetic Energy/Nucleon*

Kinetic energy/nucleon (MeV)	Integral flux in particles/(m² sr sec)
65.5	207.6±11.3
100	205.0±11.2
200	197.7±11.2
300	187.7±11.1
400	173.8±11.1
500	158.9±10.8
600	151.2±10.6

incorrectly identified as He^4, has been investigated by Waddington[44] and shown to be small, although the astrophysical significance of a large percentage of He^3 may be very great.[45] Furthermore energies determined from ionization depend only on the charge and the only error made is in the correction for energy loss to the top of the atmosphere.

In order to calculate the intensity and the differential energy spectrum, a smooth trial function which consisted of the He-particle density divided by an estimated effective exposure time was assumed. The shape of this trial function after passing through various amounts of atmosphere was constructed from range energy tables[41] and then reduced in magnitude in accordance with the combined effects of the absorption by interaction of He-particles and the production of He-particles by heavy nuclei interactions in the air. The mean free paths and fragmentation parameters for emulsion and air were those listed in the review article by Waddington.[46] The secondary helium nuclei were assumed to have the same energy spectrum as the primary particles and the effect of the different rate of energy loss of the heavy primary parents was ignored since secondary helium nuclei formed in the air and the emulsions amount to only a few percent of the total and changes in the energy spectrum from the above considerations are small and uncertain.

In principle, the total contribution to the particle density by the trial function can then be calculated by integrating the appropriate energy spectrum at a given point and angle over the solid angle of acceptance for the known orientation of the stack at that time and then integrating over the entire flight. In practice, these integrals were approximated by sums consisting of average values within periods. The ascent was broken into steps and the floating altitude was divided into the short portion when the stack was still inverted and the portion when it was upright. For particles whose energy was less than 200 MeV/nucleon, only the last portion was included since the particle direction for these low-energy particles was known and particles whose direction was opposite to the downward one at ceiling were rejected. The resulting particle density in units of particles/(cm² sr) was then compared with the observed particle density. On the basis of this result, a better estimate of the primary spectrum, or in essence the effective time, was made. The second trial function fitted the observed data. Both the particle densities and the flux values are given in Table II, and the differential spectrum is plotted in Fig. 3 along with the proton data.

COMPARISON WITH OTHER RESULTS

Before discussing the implications of the experimental results with regard to the time variations of cosmic rays and specific modulation mecha-

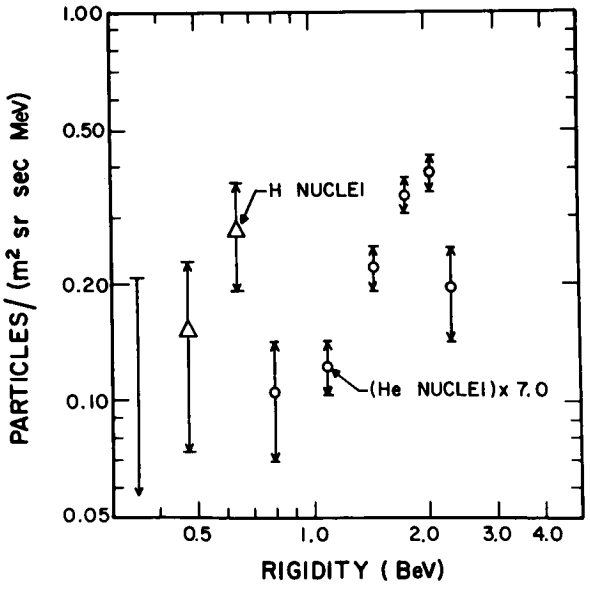


FIGURE 3.—Differential rigidity spectra for hydrogen nuclei, and helium nuclei times 7.0.

isms, the experimental results will be compared with those obtained at other times in the solar cycle. The combined picture will then give a better basis for comparison with any particular theory.

Hydrogen Nuclei

In Fig. 4 the differential energy spectrum is plotted together with the results of Vogt[18] and Meyer and Vogt[19] from balloon flights at Fort Churchill on 22 August 1960; 15 September 1960; and 8 August 1961, and the Explorer XII results of Bryant et al.[47] on 18 August 1961. The lowest energy point of Bryant et al. has a somewhat greater uncertainty than that shown because of calibration changes. This point is currently being recalculated.[48] At energies above about 150 MeV the differential proton flux values are seen to be approximately the same for each measurement. However, at lower energies there exists a marked difference between the data obtained by Meyer and Vogt and the present work. Whereas Meyer and Vogt observe a definite increase in the differential intensity at lower energy both in 1960 and 1961, the pretent data indicate a decrease. At least in the case of the 1961 flights, the

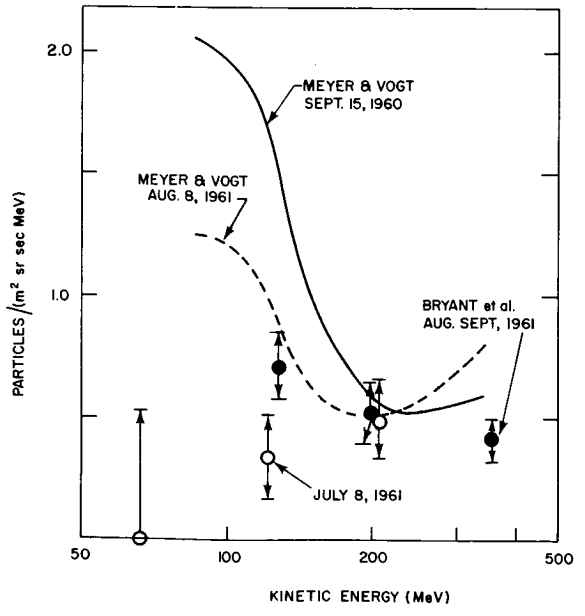


FIGURE 4.—Differential energy spectra for hydrogen nuclei measured at various times.

difference in intensity in the energy interval from approximately 80 to 160 MeV may possibly be largely explained in terms of the different corrections used for secondary production in the atmosphere above the balloon. If the method of correction used in this work is applied to the Meyer and Vogt[20] raw data, a differential flux of 0.59 ± 0.10 particles/m^2 sr sec is obtained rather than their value of 0.82 ± 0.05 particles/m^2 sr sec, whereas this work gives 0.34 ± 0.17 particles/m^2 sr sec. When the correction used by Meyer and Vogt is compared to that of this work, it is seen that the major difference is the absolute number of secondary grey tracks assumed to be produced on the average by cosmic ray particles in a nuclear emulsion and not the conversion from interactions in nuclear emulsions to those in air. Meyer and Vogt's estimate of the number of grey secondaries in air is about a factor of 1.7 smaller than the one of this work. This correction used here is based on direct measurements made on secondary cosmic-ray particles formed in the same emulsions from which the flight data was obtained. Meyer and Vogt[19] used the previous calculations of Vogt[19,49] which are based on the analysis of area scans for stars in nuclear emulsions flown at lower altitudes and the extrapolation of these results to the top of the atmosphere. A combination of factors, such as statistical fluctuations, uncertainties in extrapolation, the possible low detection efficiency for one- and two-prong stars by the area scanning method, and small differences in other factors, may combine to account for the discrepancy.

Another possible contribution to the difference in intensities is the fact that, although the flights of Meyer and Vogt considered here were not made immediately following a known injection of solar particles and are considered by them to be typical of a quiet day, they were flown at times during which the sun was quite active. The flight on 22 August 1960 was flown after a large Forbush decrease following a probable solar particle event on 11 August,[50] and the flight on 15 September 1960 was flown during the recovery phase of a Forbush decrease following the solar particle event of 3 September 1960. The flights in August 1961 were also flown during the recovery phase of a Forbush decrease and following the series of large solar particle events of 11–26 July. It might then

be said that, while these flights were not flown when known injections of solar particles were in progress, they were flown either at times when such injections were likely to occur, or when it was possible that some of the radiation from a previous solar particle event was still trapped in a region of the solar system containing the earth.

The flight reported here, on the other hand, was flown at a time that was much less disturbed as indicated in Sec. IIA. These data do not conflict, then, with the results of Meyer and Vogt, but rather, they are a measurement of the spectrum of hydrogen nuclei at a time of much lower solar activity.

The question remains whether or not the low-energy spectrum measured in this experiment represents the galactic cosmic-ray protons during this time in the solar cycle or whether there is a contribution of solar particles even during this period of a very quiet sun. The question cannot be unambiguously answered without further measurements as solar minimum is approached, if even then. If the intensity is observed to decrease, it would indicate that at least a portion, if not all, of these low-energy protons are probably of solar origin. If the intensity is observed to increase with the approach to solar minimum, it is possible, although not necessarily true, that the spectrum represents the modulated galactic cosmic-ray proton intensity.

Helium Nuclei

Webber[2] has compared the variation in the He-particle intensity above 1.5 BV, 260 MeV/nucleon, with the Mt. Washington neutron monitor rate during the period when the cosmic-ray intensity declined from its maximum value to its lowest values, specifically the years 1954 to 1959, and shown that there is a unique relation between these two parameters. As Webber has noted, the total He flux varies at a rate of 2.1 to 3 times as fast as the neutron monitor counting rate, as expected since the neutron monitor rate reflects primarily the variations in the flux of high-rigidity particles, which vary less than the flux of low-rigidity particles. Webber's smooth curve is shown in Fig. 5 along with the experimental result of this work. This figure shows that the experimental data of this work were obtained at a time between maximum and minimum intensity, and that the flux observed falls on this smooth curve which was determined from data taken during the decline of the cosmic-ray flux. A similar analysis by Stevenson and Waddington[15] of the relation between the α-particle, $J_\alpha{}^0$ above 200 MeV/N and the Ottawa neutron monitor hourly rate N showed that the data were best fitted by a quadratic of the form.

$$J_\alpha{}^0 - 189 = -(13.4 \pm 9.9) + (N-285)(2.48 \pm 0.41) + (N-285)^2(0.043 \pm 0.0024).$$

At the time of this flight N was 291.0 counts/h (scaling factor 64), which predicts $J_\alpha{}^0 = 192 \pm 10$ particles/cm^2 sr sec and compares well with the value of 198 ± 11 found in this experiment. Therefore, since the neutron monitor rate is a measure of the high-energy intensity, this result suggests that during the period of increasing cosmic-ray intensity the ratio of low-rigidity particles to high-rigidity ones at a given intensity is the same as during the declining phase.

Having examined the coarse features of the variation, the detailed differential energy spectrum is now compared with those observed at solar

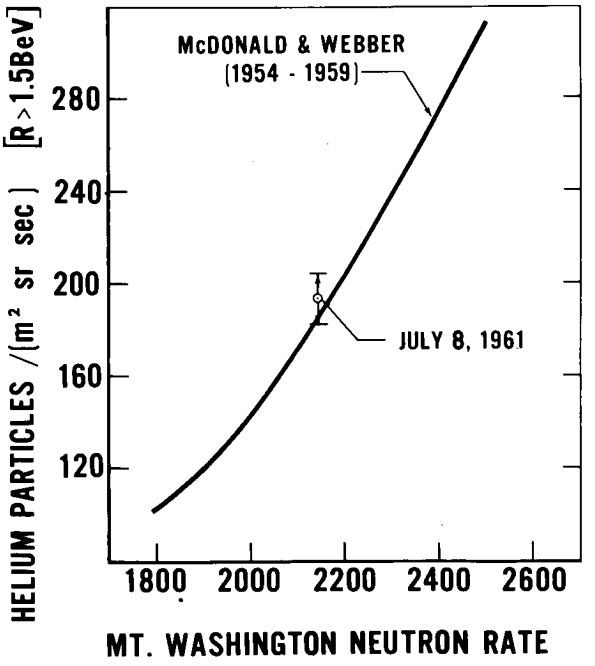


FIGURE 5.—Counting rate of Mount Washington neutron monitor vs. helium nuclei flux.

maximum and minimum. The results of experiments during these two periods are shown in Fig. 6(a). From these data smooth curves are then drawn in Fig. 6(b) and compared with the results of the present experiment. The values of this experiment are seen to lie between the spectra typical of solar minimum and maximum. In addition, the peak in the spectrum appears to have moved to a rigidity intermediate between the peak rigidities during the other two periods.

This result, together with the earlier figures showing helium particle intensity as a function of neutron monitor counting rate, supports the hypothesis that the recovery of the differential intensity is continuous over the whole rigidity range, that is, there is no tendency for the high- or low-rigidity particles to recover preferentially.

Modulation of Cosmic Rays

A large number of theories attempting to describe the modulation of cosmic rays have been developed in recent years. We shall discuss here only those known to us which have been developed quantitatively and for which there is at least partial agreement with experimental results. A summary of modulation models has recently been made by Webber,[2] and the reader is referred to this article and the references contained therein for a more complete discussion.

Nagashima[51] originally suggested that the modulation of the primary cosmic radiation could be explained by the positive cosmic-ray particles having to pass through a geocentric decelerating static electric potential, which then could vary throughout the solar cycle, while Ehmert[52] has considered the effect of a heliocentric potential. By Liouville's theorem, the quantity $(dj/dP)/p^2\beta$, and hence $(dj/dW)/p^2$, must remain constant for a set of particles. Here, j is the particle flux, p the particle momentum, and β the particle velocity. Hence, for a decelerating potential V, the following expression can be obtained for a particle of charge Ze and atomic number A:

$$\left.\frac{dJ}{dW'}\right]_{[W-ZeV/A]} = \left.\frac{dJ}{dW'}\right]_{[W]} \left[\frac{(W-ZeV/A)^2-M_0C^2}{W^2-M_0C^2}\right],$$

where W is the total energy per nucleon before deceleration and M_0C^2 is the rest mass per nucleon. Hence, if the energy of the particle decreases, dj/dW will decrease also by an amount which increases with decreasing energy. McDonald and Webber[3] and Fichtel[36] have shown previously that reasonable agreement can be obtained with experimental data for helium and heavy nuclei. Ehmert[52] has shown that if a potential of about 1 BeV exists at sunspot minimum, the peak in the differential spectrum even at that time can be explained.

The presence of a quasistatic electric field sufficiently large to cause the observed reduction in intensity seems inconsistent with the present estimates of the physical conditions existing in the solar system, particularly the high conductivity, and the low abundance or absence of energetic electrons which would have been accelerated by the electric field. In addition, McDonald[4] and McDonald and Webber[53] have shown that to within experimental errors, when measurements were made on the helium and proton components at about 1.3-BV rigidity the intensities were reduced by the same proportion between solar minimum and maximum. For a given rigidity, the electric deceleration model predicts that helium nuclei should be more suppressed than protons because their charge-to-mass ratio is half that of the protons.

In this experiment an extrapolation of the α-particle spectrum would suggest a splitting in the rigidity spectra of the proton and helium

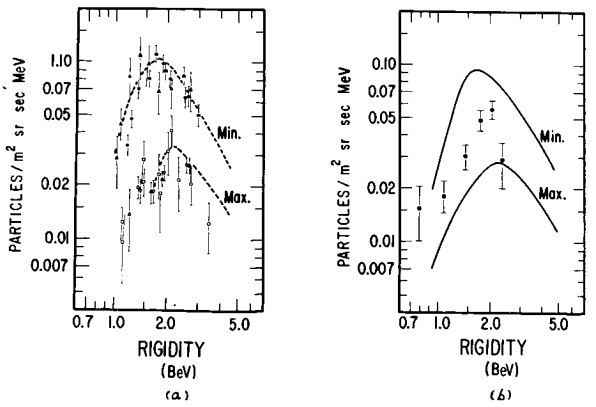


FIGURE 6.—(a) Differential spectra of helium nuclei at solar maximum and solar minimum. ▲ reference 13, ■ reference 4, ● reference 4, ▽ reference 4, △ reference 4, ○ reference 10, □ reference 8, --- reference 6; (b) Data from the flight of 8 July 1961, compared with curves 6(a).

nuclei components in a manner consistent with a decelerating model; however, in view of the physical implausibility of such a model, and the previous work of McDonald and Webber it seems reasonable to look for another explanation. Obviously the presence of additional solar protons would provide an explanation for the observed splitting.

Parker[54,55] has proposed a diffusion model, wherein the solar wind, a flow of gas consisting of a distorted magnetized plasma whose existence was first deduced by Biermann[56,57] is considered to be responsible for the modulation. In this theory, an equilibrium state is established wherein both the effects of diffusion through the shell and removal by convection are considered. He obtains an equation of the form

$$dJ/dR = (dJ/dR_{(\infty)})[\exp - \{K(t)/(\beta\lambda)\}],$$

where λ is proportional to R^2 if the average dimension of the scattering clouds is much smaller than the radius of curvature and is a constant if the average dimension is much larger than the radius of curvature.

As in the case of the electric deceleration model and other diffusion models, the modulation at a given rigidity is different for particles with different charge to mass ratios because β appears, and therefore, the comments made previously concerning relevant experimental results apply here also. Further, it seems difficult to find a reasonable choice of parameters and a reasonable dJ/dR_∞ which will yield an expression which agrees with both the helium and the hydrogen nuclei spectral data for λ either a constant or proportional to R^2.

The modulation of the cosmic radiation by a dipole field of the sun was first proposed by Janossy.[58] In order to be effective the dipole moment of the sun needs to be about 10^{34} G cm^3, and hence a polar field of about 30 G should exist. The actual general solar field has an intensity of about 1 G at the poles[59] and, in fact, seems to go through zero at sunspot maximum[60] when the depression of the cosmic-ray intensity is greatest.

Elliott[61] tried to overcome this difficulty by suggesting that large current systems exist in the corona at a distance between 5 and 15 solar radii from the sun and that these ring currents should produce a dipole field at the earth's orbit and beyond. This ordered field should often be disturbed by plasma streams from the sun. These perturbations scatter particles into orbits which then enter a disordered field region around the sun, resulting in absorption of particles which reduces the intensity in the forbidden regions. This theory predicts a rigidity dependence of the modulation which can be brought into satisfactory agreement with experimental rigidity spectra by suitable adjustment of the parameters. There are, however, at least two experimental measurements which are apparently in confluct with this model. Firstly, measurement of the magnetic field at distances up to 32 earth radii from the earth by Explorer X[62] show that during "quiet" periods the field direction is apparently not that of a dipole. Secondly, the lack of any detectable variation in the radiation between the earth and Venus, as measured on Mariner II,[62] is inconsistent with the variation of about 30% predicted by Elliot.[63]

SUMMARY

The helium-particle data shows that during the recovery phase of the cosmic-ray modulation cycle the differential helium rigidity spectrum lies in between that at solar maximum and solar minimum. Further, a comparison of the relation of the integral helium flux and neutron monitor counting rate with those during the declining phase of cosmic-ray intensity, namely, from solar minimum to solar maximum, indicated that the functional dependence was the same. Therefore, there was good reason to believe that the modulation mechanism produced the same rigidity dependence both during the declining and recovery phase.

Because there is apparently a decrease in the proton flux in the vicinity of 0.5-BV rigidity with decreasing solar activity, whereas the proton flux above 1.0 BV shows an inverse correlation with solar activity, it seems likely that there was a non-negligible solar proton component at 0.5-BV rigidity at least during the two flights of Vogt[19] in 1960. Considering the proton and helium nuclei data of this work together, it seems that either there is a modulation mechanism where in the relative depression of the protons with respect to the helium nuclei is nearly the same at 1.3-BV

rigidity, but is markedly different at 0.5 BV, or there is a small solar proton component. The possibility of a significant solar contribution seems less likely because the proton differential flux in this work shows no increase with decreasing energy whereas a leveling off and an increase might be expected if there were an appreciable solar component. Further, none of the low-energy differential flux values of Meyer and Vogt[20] are below those of this experiment, even if the correction used in this work is applied to their data. Data from subsequent years should help to answer this question.

REFERENCES

1. S. E. Forbush, J. Geophys. Res. **59,** 557 (1954).
2. W. R. Webber, *Progress in Elementary Particle and Cosmic Ray Physics* (Interscience Publishers, Inc., New York, 1962), Vol. VI.
3. F. B. McDonald and W. R. Webber, Phys. Rev. **115,** 194(1959).
4. F. B. McDonald, Phys. Rev. **116,** 462 (1959).
5. F. B. McDonald and W. R. Webber, J. Geophys. Res. **65,** 767 (1960).
6. F. B. McDonald, J. Phys. Soc. Japan Suppl. A–II, **17,** 428 (1962).
7. J. J. Quenby and C. J. Wenk, Phil. Mag. **7,** 1457 (1962).
8. H. Aizu, Y. Fujimoto, S. Nasegawa, M. Koshiba, I. Mito, J. Mishimura, K. Yokai, and M. Shein, Phys. Rev. **116,** 436 (1959).
9. P. J. Duke, Phil. Mag. **5,** 1151 (1960).
10. A. Engler, F. Foster, T. L. Green, and J. Mulvey, Nuovo Cimento **20,** 1157 (1961).
11. A. Engler, M. F. Kaplan, J. Klarmann, A. Kennan, C. Fichtel, and M. W. Friedlander, Nuovo Cimento **19,** 1090 (1961).
12. P. H. Fowler, P. S. Freier, and E. P. Ney, Nuovo Cimento Suppl. **8,** 492 (1958).
13. P. S. Freier, E. P. Ney, and C. J. Waddington, Phys. Rev. **114,** 365 (1959).
14. G. R. Stevenson, Nuovo Cimento **24,** 557 (1962).
15. G. R. Stevenson and C. J. Waddington, Phil. Mag. **6,** 517 (1961).
16. C. J. Waddington, Phil. Mag. **45,** 1312 (1954).
17. J. R. Winckler and K. A. Anderson, Phys. Rev. **108,** 48 (1957).
18. H. V. Neher, Phys. Rev. **107,** 588 (1957).
19. R. Vogt, Phys. Rev. **125,** 366 (1962).
20. P. Meyer and R. Vogt, Phys. Rev. **129,** 2275 (1963)
21. The data from which this absorption curve was deduced were provided through the courtesy of Dr. T. R. Hartz and Dr. E. L. Vogan of the Canadian Defense Research Telecommunications Establishment.
22. H. Carmichael (private communication).
23. J. V. Lincoln, J. Geophys. Rev. **66,** 3949 (1961).
24. P. H. Fowler and D. H. Perkins, Phil. Mag. **46,** 587 (1955).
25. A grey track is conventionally defined as one with an ionization that is greater than 1.4 or 1.5, and less than 6 times the proton minimum.
26. N. Metropolis, R. Bivins, M. Storm, A. Turkevich, J. M. Miller, and G. Friedlander, Phys. Rev. **110,** 185 (1958).
27. N. Metropolis, R. Bivins, M. Storm J. M. Miller, G. Friedlander, and A. Turkevich, Phys. Rev. **110,** 204 (1958).
28. H. W. Bertini (private communication).
29. N_h is conventionally defined as the number of secondary tracks from an interaction which have an ionization ≥ 1.4 times the proton minimum.
30. W. O. Lock, P. V. March, and R. McKeague, Proc. Roy. Soc. (London) **231,** 368 (1955).
31. W. W. Brown, Phys. Rev. **93,** 528 (1954).
32. V. Beliakov, Van Shu-fen', V. Glagolev, Dalkhazhav, L. Kirillova, P. Markov, R. Lebedev, K. Tolstov, E. Tsyganov, M. Shafranova, Jao Tsyng-se, B. Bannik, G. Bajatjan, I. Gramenitskij, M. Danysz, N. Kostanashvili, V. Lyubimov, A. Nomofilov, M. Podgoretskij, E. Skshipchak, D. Tuvdendorge, O. Shahulashvili, N. Bogachev, S. Bunyatov, T. Vishki, Yu-Merekov, and V. Sidorov, Proc. 1958 Annual Conf. on High Energy Phys. at CERN, 309 (1958).
33. G. B. Zhdanov, V. M. Maksinenko, M. I. Tret'Yakova, and M. N. Shcherbakova, Zh. Eksperim. i Teor. Fiz. **37,** 620 (1959) [translation: Soviet Phys.—JETP **10,** 442 (1960)].
34. C. J. Waddington, Phil. Mag. **5,** 311 (1960).
35. C. J. Waddington, Nuovo Cimento Suppl. **19,** 37 (1961).
36. C. E. Fichtel, Nuovo Cimento **12,** Suppl. **19,** 1100 (1961).
37. P. B. Fowler, Phil. Mag. **41,** 169 (1950).
38. M. G. K. Menon, C. O'Ceallaigh, and O. Rochat, Phil. Mag. **49,** 932 (1951).
39. C. E. Fichtel and M. W. Friedlander, Nuovo Cimento **10,** 1032 (1958).
40. P. S. Freier, E. P. Ney, and C. J. Waddington, Phys. Rev. **114,** 365 (1959).
41. J. H. Atkinson, Jr., and B. H. Willis, University of California Laboratory Report No. UCRL 2426, Revision II (unpublished).
42. B. Hildebrand, F. W. O'Dell, M. M. Shapiro, R. Silberberg, and B. Stiller, Bull. Am. Phys. Soc. **7,** 311 (1962).
43. F. Foster and J. H. Mulvey, Nuovo Cimento **27,** 93 (1963).
44. C. J. Waddington, Proc. of XIXth Varenna Summer School (1961).
45. M. V. K. Appa Rao and M. F. Kaplon, Nuovo Cimento **21,** 369 (1961).
46. C. J. Waddington, Progr. Nucl. Phys. **8,** 3 (1960).
47. D. A. Bryant, T. L. Cline, V. D. Desai, and F. B. McDonald, J. Geophys. Res. **67,** 4983 (1962).

48. F. B. McDonald (private communication).
49. R. Vogt, thesis, The University of Chicago, 1961 (unpublished).
50. J. B. Gregory, J. Geophys. Res. **68,** 3097 (1963).
51. K. Nagashima, J. Geomag. and Geoelec. **3,** 100 (1951).
52. A. Ehmert, Proc. Moscow Cosmic Ray Conf. IV, 142 (1960).
53. F. B. McDonald and W. R. Webber, *Proceedings of the First International Space Science Symposium, Nice* (North-Holland Publishing Company, Amsterdam 1960), p. 968.
54. E. N. Parker, Phys. Rev. **109,** 1874 (1958).
55. E. N. Parker, Phys. Rev. **110,** 1445 (1958).
56. L. Biermann, Z. Astrophys. **29,** 274 (1951).
57. L. Biermann, Observatory **77,** 109 (1957).
58. L. Janossy, Z. Physik **104,** 430 (1937).
59. H. D. Babcock, Astrophys. J. **130,** 364 (1959).
60. H. D. Babcock, Astrophys. J. **133,** 572 (1961).
61. H. Elliot, Mag. **5,** 601 (1960).
62. H. R. Anderson, Science **139,** 42 (1963).
63. H. Elliot, Nature **186,** 299 (1960).

SEPTEMBER 26, 1960 SOLAR COSMIC RAY EVENT*

C. E. FICHTEL, D. A. KNIFFEN

AND

K. W. OGILVIE

Goddard Space Flight Center

Type IV radio emission is thought to be the most important observable characteristic of solar flares producing high energy particles. We show here that solar cosmic ray particles arrived at the earth on September 27, 1960 following a small flare accompanied by a type IV burst of short duration. Apart from their magnitude, the characteristics of this event do not differ essentially from those of larger events.

INTRODUCTION

Considerable attention has been directed toward the study of major solar cosmic ray events, especially during the present solar cycle.[1] These major events have been characterized by integral fluxes of protons above 80 Mev, which were of the order of 100 times the cosmic ray background, and by large, easily observable, cosmic noise absorption—"polar black-outs"—as detected by a ground-level riometer. Increases in the counting rates of sea-level neutron monitors indicate that in some events the proton energy spectrum has a large high energy component. It is known that essentially all major solar cosmic ray events follow major flares which are associated with type IV radio bursts,[2,3,4] but not all type IV radio emissions have been followed by major solar cosmic ray events. It has been suggested by Hakura and others[5,6,7] that type IV radio emissions are due to the synchrotron radiation of electrons accelerated in solar events in which protons were also being accelerated. This naturally suggests the possibility that solar protons, presumably in much smaller numbers than in the large events, also follow small type IV radio bursts.

To assist in determining whether or not type IV radio emissions are related to solar cosmic ray events in general, and to study the properties of any particles that might be present in a smaller event, a sounding rocket carrying nuclear emulsions and counter instruments was fired on September 27, 1960, at 1444 UT, from Churchill, Canada, into an event of the aforementioned smaller type.

DESCRIPTION OF THE EVENT

Table 1 lists the flares that were observed during the period September 24–September 27, 1960. One of these flares—beginning at 0525 UT and ending at 0605 UT on September 26, 1960—is particularly interesting since it is the only one of the group which had an associated type IV radio burst. It was reported by CSIRO (Sydney), which also reported type III and type II bursts at about the same time. The time interval during which the type IV continuum was observed was relatively short, 0554 to 0611 UT September 26, 1960.

The 0525 UT flare had several other characteristics which have been shown to be related to

*This report has been published in substantially the same form in *J. Geophys. Res.* **67**(10): 3669–3672, September, 1962.

TABLE 1.—*Solar Flares September 24 to 27, 1960 (Importance 1+ or Above)*

Date (1960)	Beginning of flare (UT)	Class of importance	Plage number	Approximate solar coordinates (degrees)	Ionospheric effect
September 24	0714	1 to 1+	5863	S21 W01	
	0923	1 to 1+	5863	S20 E01	
	2116	1 to 2+	5863	S20 W10	
	2304	1+	5866	N26 E19	
September 25	0759	3	5866	N26 E09	
	1246	1 to 2	5863	S18 W10	
	1932	2− to 2	5866	N27 E06	
September 26	0525	*1 to 2+	5858	S21 W64	Slow S–SWF
	1350	1 to 2	5858	S21 W63	Slow S–SWF
September 27	1250	1 to 2−	5863	S13 W24	

There were no flare patrol observations during the following intervals:
September 24: 0245 UT–0300 UT
0515 UT–0600 UT
1315 UT–1330 UT
September 25: 0200 UT–0345 UT
0415 UT–0500 UT
September 27: 0130 UT–0500 UT

*Classified by Tashkent as 2+ and by Mitaka as 1.

solar cosmic ray events: (1) There was a S–SWF (Sudden Short Wave Fadeout) of intensity 2+. (2) The plage group in which the flare was located was reasonably well developed and had one penumbral area which was large enough to classify it, according to the Anderson criterion,[8] as a region likely to produce a solar cosmic ray event. (3) The same plage group had been seen on a previous rotation of the sun. Dodson[9] has shown that the second rotation is the most probable one for the occurrence of a major polar cap absorption event.

The small amount of absorption detected by riometers on September 26, 27 and 28, 1960[10,11] has features characteristic of polar cap events. The absorption began at Churchill about an hour after the type IV radio burst and was generally greater during the day than at night, always remaining greater than zero. However, the sunrise onset and sunset recovery was not uniformly clear at all stations.

EXPERIMENTAL RESULTS

Since a discussion of the Nike–Cajun payload system employed in the Solar Beam Experiment, the flight characteristics of the rocket, and the general method of reducing the data derived from the nuclear emulsions and particle detectors are presented fully in the papers by Davis,[12] Biswas[13] and Ogilvie,[14] only a brief statement will be given here. The nuclear emulsions are exposed to the ambient radiation for an equivalent period of about 135 seconds.[12] The corrections to the emulsion results for ascent, descent, background, measurement uncertainties and attitude of the rocket are relatively small[10,11,12] and can be made with a combined uncertainty of about 10 percent. Therefore, a large fraction of the error in individual data points is statistical. The proton energy spectrum obtained is shown in figure 1. The low energy cutoff of the kinetic energy at 13.5 Mev is determined by the material between

the depth in the nuclear emulsions at which the energy measurements were made and the ambient radiation, since the threshold due to the magnetic field is much lower than this at Churchill. The high energy limit is set by the limited collecting time and the small flux of particles. For purposes of comparison in the low energy region, essentially identical firings were made from Churchill, on June 6, 1960 and December 10, 1961, during geophysically quiet times. The September 27 proton flux in the 13–50 Mev interval is at least twenty-five times normal.

We now turn to a discussion of the counter instrumentation results. An unfortunate power failure in the telemetry recorders during the flight of September 27 has prevented a full analysis of the counter observations. It is, however, possible to present some information, which is, of necessity, of low accuracy. In figure 1, the solid triangle at 30 Mev represents the results of the geiger counter measurement. It must be emphasized that these results, in contrast to the usual practice, are not independent of the emulsion results. The slope of the energy spectrum measured by means of the emulsions has been used to compute corrections in the range for which it is applicable. The points shown at 4.5, 2.8 and 1.8 Mev represent the results of the scintillator measurements. We were unable to employ our usual method of ratios, but the fortunate circumstance of the steep spectrum allows approximations to be made. Their effect is reflected in the large errors, but the results demonstrate that the spectrum undoubtedly curves over at low energies in the region of 10 Mev. If the spectrum at higher energies had persisted down to a few Mev the rates observed by the scintillators would have been at least an order of magnitude higher than observed.

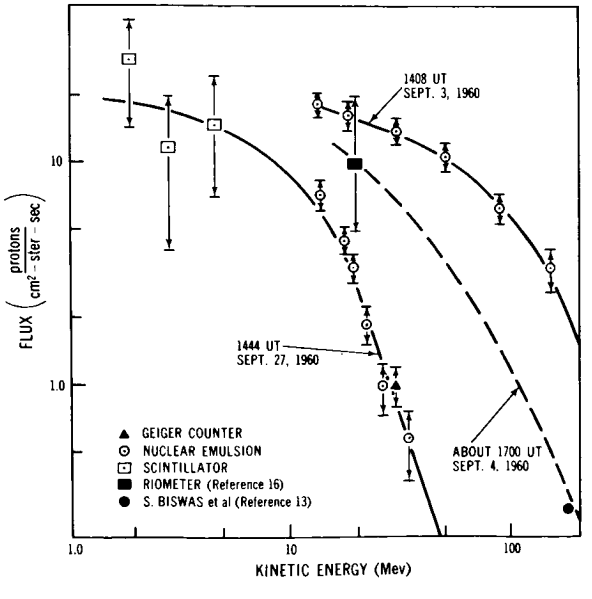


FIGURE 1.—Primary proton energy spectrum.

Later measurements, made at the end of the event of November 15, 1960 when the magnetic field had returned to its normal value, were consistent with a threshold of 4.5±0.5 Mev at Churchill.[14]

On the 27th of September 1960, the magnetic conditions immediately before the rocket firing were fairly disturbed.[15] There was a magnetic storm with a small initial impulse which began at 1930 UT on the 26th and ended at 1100 UT on the 27th. The magnetic threshold was, therefore, 4.5 Mev or less at the time. Thus we may say with certainty that the integral spectrum curves over in an energy region well above the threshold.

If a curve of the form $dj/de = N_0/E^{\gamma(E)}$ is fitted to the data points, γ is found to be 3.6±0.6 at 25 Mev. The slopes of the spectra found in other events are presented in Table 2. Since this quantity varies with time we have quoted results of observations taken as close to the same time after the corresponding flare as possible. The slope of the September 27, 1960 differential proton spectrum at 25 Mev, about 33 hours after the flare, is seen to be somewhat steeper than that of the November 12, 1960 or the November 15, 1960 proton spectra at comparable times. The slope of the September 3, 1960 proton spectrum is seen to be fairly small relatively early in that event. Although no data was available from this laboratory late in the event, an approximate integral energy spectrum about forty hours after the September 3, 1960 flare was constructed by using the results of Biswas[13] at high energies and the estimated flux at 20 Mev based on riometer data.[16] The detailed energy spectra for these flights are compared in figure 1.

The fact that the energy spectrum of the September 27, 1960 event is steeper than the other three measured should not be taken to mean

TABLE 2.—*Slope of Differential Energy Spectra for E=25 Mev*

Event	Time from beginning of associated flare (hours)	Solar coordinates (degrees)	Slope
September 3, 1960	14	N20 E87	*0.6±0.3
September 26, 1960	33	S21 W64	3.7±0.6
November 12, 1960	27	N26 W5	2.1±0.3
November 15, 1960	42	N26 W35	2.1±0.3

*A value of 1.3 is obtained 40 hours after the flare by the method discussed herein, but this number is very uncertain.

that this event is different from all others. For example, consider the August 29, 1957 event whose energy spectrum late in the event was also very steep. The flux at 20 Mev as deduced from riometer absorption was a few hundred particles/cm-ster-sec, yet Anderson[17] saw no detectable increase at balloon altitudes where the energy threshold was 100 Mev. The slope of the differential energy spectrum of this latter event between 20 and 100 Mev must then have been of the order of 6 or more.

CONCLUSION

We have shown that, in at least one case, solar cosmic ray particles arrived at the earth following a relatively small flare, but one accompanied by a short duration type IV radio burst. Further, the September 26, 1960 event fits well into what seems to be a developing pattern for solar cosmic ray events. These events normally follow flares accompanied by type IV radio emission; their proton spectra steepen at high energies and have a slope of the order of 1 to 6 in the ranges of tens of Mev late in the event. These general features will be elucidated by means of the additional data which will be forthcoming from the experiments of Davis, McDonald, Van Allen and others on Explorer XII, which was launched in August of 1961.

ACKNOWLEDGMENTS

We wish to express our thanks to Dr. H. Dodson of the McMath-Hulbert Observatory and to Mrs. H. H. Malitson of the Goddard Space Flight Center, for interesting discussions of the event and for supplying us with solar activity information related to the September 26, 1960 period. During the time of this firing, the field operations were under the able direction of Mr. L. R. Davis.

REFERENCES

1. WEBBER, W. R., "Time Variations of Low Energy Cosmic Rays during the Recent Solar Cycle," *Progress in Elementary Particle and Cosmic Ray Physics*, ed. by J. G. Wilson and S. A. Wouthuysen, Amsterdam: North-Holland Publ. Co., Vol. 6, 1962 (in press).
2. HAKURA, Y., and GOH, T., "Pre-SC Polar Cap Ionospheric Blackout and Type IV Solar Radio Outburst," *J. Radio Res. Lab.* (Japan), **6**(28): 635–650, October 1959.
3. OBAYASHI, T., and HAKURA, Y., "Enhanced Ionization in the Polar Ionosphere Caused by Solar Corpuscular Emissions," *Rept. Ionosphere and Space Res. in Japan*, **14**(1): 1–40, March 1960.
4. THOMPSON, A. R., and MAXWELL, A., "Solar Radio Bursts and Low-Energy Cosmic-Rays," *Nature*, **185**(4706): 89–90, January 9, 1960.
5. BOISCHOT, A., and DENISSE, J. F., "Radio Astronomic Les Emissions de Type IV et L'origione des Rayons Cosmique Associes aux Eruptions Chromospheriques," *Compt. Rend.*, **245**, pt. II: 2194–2197, December 16, 1957.
6. ANDERSON, K. A., ARNOLDY, R., et al., "Observation at Low-Energies of Solar Cosmic Rays from the Flare of 22 August 1958, *J. Geophys. Res.*, **64**(9): 1133–1147, September 1959.
7. HAKURA, Y., "Some Statistics on the Solar Cosmic Rays Produced by Solar Eruptions Associated with Type IV Outbursts," Private Communication, 1961.

8. ANDERSON, K. A., "Preliminary Study of Prediction Aspects of Solar Cosmic Ray Events," NASA Technical Note D-700, April 1961.
9. DODSON, H., Private Communication, 1961.
10. LEINBACH, H., Private Communication, 1961.
11. HARTZ, T. R., Private Communication, 1961.
12. DAVIS, L. R., FICHTEL, C. E., et al., "Rocket Observations of Solar Protons on September 3, 1960," *Phys. Rev. Letters*, **6**(9): 492–494, May 1, 1961.
13. BISWAS, S., FREIER, P. S., and STEIN, W., "Solar Protons and α Particles from the September 3, 1960 Flares," *J. Geophys. Res.*, **67**(1): 13–24, January 1962.
14. OGILVIE, K. W., BRYANT, D. A., and DAVIS, L. R., Rocket Observations of Solar Protons during the November 1960 Events, 1," *J. Geophys. Res.*, **67**(3): 929–937, March 1962.
15. LINCOLN, J. V., "Geomagnetic and Solar Data," *J. Geophys. Res.*, **66**(1): 311–315, January 1961.
16. FICHTEL, C. E., GUSS, D. E., and OGILVIE, K. W., "Details of Individual Solar Particle Events," *Solar Proton Manual*, NASA Technical Report, 1962.
17. ANDERSON, K. A., and ENEMARK, D. C., "Observations of Solar Cosmic Rays Near the North Magnetic Pole," *J. Geophys. Res.*, **65**(9): 2665, September 1960.

OBSERVATIONS ON THE SOLAR PARTICLE EVENTS OF JULY 1961

D. E. GUSS AND C. J. WADDINGTON*

Goddard Space Flight Center

The spectrum of solar protons with energies between 77.5 and 290 Mev has been measured during the flare event of July 12, 1961 with nuclear emulsions flown on a high altitude balloon from Fort Churchill. The differential spectra averaged over the period from 0839 UT to 1819 UT on July 13 are consistent with a power law of kinetic energy with an exponent $\alpha = 5.6 \pm 0.3$, a power law of total energy over rest energy with $\gamma = 47 \pm 3$, and an exponential law of rigidity with $P_0 = 52 \pm 3$. The intensity of particles with $E \geq 77.5$ Mev $= 1.69 \pm 0.14$ p/cm² sr sec. During the solar particle event of July 18 the intensity of particles with $E \geq 100$ Mev was 40 ± 8 p/cm² sr sec in the time interval from 1305 UT to 1918 UT on July 18. The similarity between these events andthose of April 29 and May 4, 1960 is noted.

INTRODUCTION

In July 1961, three high altitude balloons carrying nuclear emulsions were flown from Fort Churchill, Canada. Both the second and third of these flights were made while large intensities of energetic particles associated with solar flares were arriving at the earth. In this paper we present the results of an analysis of the data obtained from these latter two flights on the proton intensities and spectra.

A chronological history of some of the more relevant data is given in figure 1. This figure shows, reading from the top: (a) the absorption of cosmic noise in db recorded by the Churchill 30 mc/s riometer†; (b) the heliographic longitudes of the major solar flares from McMath plage regions 6171 and 6172, together with an indication of their magnitudes (Natl. Bu. Standards U. S., 1961); (c) the geomagnetic planetary 3-hour range indices, Kp,‡ together with the times of occurrence of those sudden commencements (s.c.) which were observed by many stations (Lincoln, 1961, 1962); (d) the duration of each flight from time of stack flip to cut-down; and (e) the counting rate of the Deep River neutron monitor (Carmichael, 1961).

From figure 1 it can be seen that flight 1 was made during a time of quiet geophysical conditions. The results from this flight were used to correct for galactic cosmic ray background in the latter flights. During flights 2 and 3, on the other hand, conditions were extremely disturbed, and major PCA events were in progress. Furthermore, during flight 3 an increase was recorded by sea level neutron monitors. The emulsions on this latter flight were essentially saturated, and have not provided detailed data.

OBSERVATIONS ON JULY 13, 1961 (Flight 2)

Balloon Flight

A stack of twenty-three 20 cm×10 cm×0.06 cm Ilford emulsions was flown on flight 2 under

*National Academy of Sciences, NASA Senior Postdoctoral Resident Research Associate. Now at School of Physics, University of Minnesota.

†The data from which this absorption curve was deduced was provided through the courtesy of Dr. T. R. Hartz and Dr. E. L. Vogan of the Canadian Defense Research Telecommunications Establishment.

‡Notice Kp is a semi-logarithmic quantity and thus such a plot visually underemphasizes the changes in magnitude of magnetic disturbance.

0.1 gm/cm² of packing material. The stack was orientated so that its short edge was vertical and was rotated through 180° at ceiling altitude. The time-altitude curve is shown in figure 2 and shows that the most of the flight from the time the stack was flipped to cut-down was spent under 4.1 gm of atmosphere. The residual atmospheric pressure was measured to within ±0.1 gm/cm² by photographing a Wallace and Tiernan (0–20 mm of Hg) gauge.

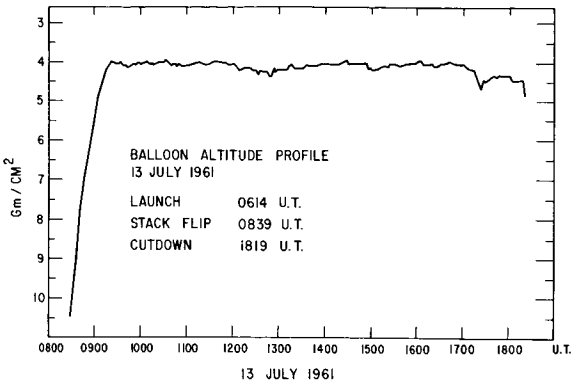


FIGURE 2.—Balloon flight profile, flight 2.

Proton Energy Spectra

From Ending Particles

As a preliminary determination of the energy spectrum of the solar protons the density of ending particles was measured at several levels in the emulsions. The observed values were corrected for ending particles produced at other times and produced by the normal cosmic ray flux by subtracting the density of particles observed in flight 1 after allowing for the differing durations of the flights.

If the stack is assumed to be infinitely thick, so that the total material traversed at any zenith angle θ at a vertical depth below the top of the atmosphere R_1 is given by $R_1 \sec\theta$, and, if the primary particles obey a differential energy spectrum of the form $dN = CE^{-\gamma}\, dE$, then the density of particles at a depth of R_1, ΔN_1, is related to the density at some other depth R_2 by (cf. Appendix I)

$$\Delta N_1/\Delta N_2 = (R_2/R_1)^{n(\gamma-1)}$$

where n is the index of the range-energy relation. Then, if the stack is assumed to have been exposed under the equivalent of 16 mm of emulsion, $\gamma = 6.3 \pm 0.4$ for $R_1 = 20$ mm, $R_2 = 30$ mms, and only 3.5 ± 0.5 for $R_1 = 30$ mm, $R_2 = 40$ mm. Obviously in this second case, the assumption that the stack is infinitely thick was not justified, and an appreciable fraction of the particles reaching the emulsion level arrive from the sides and had lower primary energies than is assumed.

It should be appreciated that this method of analysis can only be valid, even for thick stacks, when the energy spectrum is steep, so that the majority of particles arrive from directions near the zenith, as otherwise it is not permissable to neglect the effects of the nuclear absorption of particles.

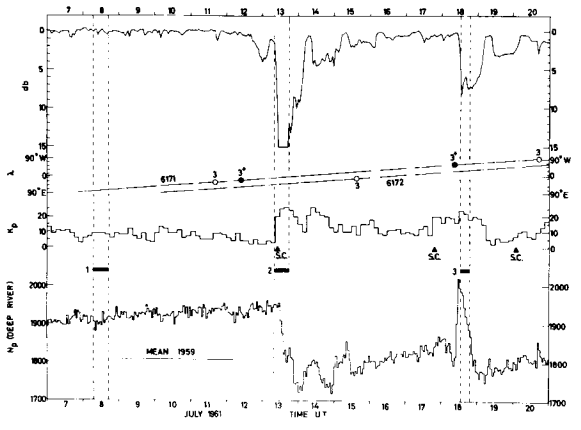


FIGURE 1.—Solar and geophysical data during July 1961; (a) absorption of 30 mc/s riometer, Fort Churchill; (b) solar flares; (c) magnetic Kp index; (d) sudden commencements; (e) balloon flights; (f) neutron minitor counting rate, Deep River.

Integral and Differential Scans

In order to make a more detailed analysis of the solar protons, integral scans were made at depths corresponding to proton energies at the top of the atmosphere of 77.5 Mev and 98 Mev while differential scans were made at depths corresponding to 90 Mev and 110 Mev. In these latter, the energy of each particle was determined either from its range or its ionization. In order to relate the differential intensities with those from the integral scans, the very small contribution from

protons with $E>290$ Mev was deduced by an extrapolation of the differential energy spectrum.

The intensities obtained from these scans were corrected for background, absorption in the overlying emulsion and atmosphere, and variations in altitude. The background correction was deduced from the energy spectra of primary and secondary particles observed in flight 1. This contribution was 1% or less for the integral scans and for the energy intervals less than 120 Mev, but rose to 50% in the interval 200 Mev to 290 Mev. The correction for absorption was made

TABLE I.—*Integral Proton Intensities for Flight 2*

Proton energy (Mev)	Integral intensity P/cm² sr sec
77.5*	1.693±0.14
90	0.856±0.065
98*	0.600±0.049
100	0.554±0.050
110	0.338±0.036
120	0.219±0.030
140	0.109±0.022
200	0.019±0.010

*Independent integral points.

using absorption mean free paths of 40 cm in emulsion and 100 gm/cm$_2$ in air. The correction for variation in flight altitude ranged from 4% at 77.5 Mev to 1% in the interval from 200 Mev to 290 Mev. Because of the steep spectrum, no correction was necessary for the period before the stack was rotated other than the background correction noted above.

The resulting intensities are shown in Table I. The integral intensity above 290 Mev was not measured, but was calculated from an extrapolation of the differential energy spectrum. The errors quoted are statistical plus systematic errors. The systematic error is about 5% and arises primarily from the uncertainties in the acceptance solid angle and collecting area.

The resulting differential energy spectra are shown in figure 3. In the range of energies studied it is not possible to distinguish between a power law of kinetic energy, a power law of total energy,

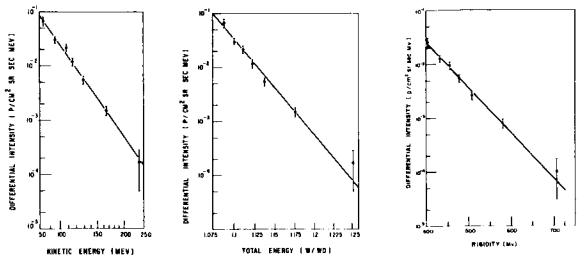


FIGURE 3.—Differential proton intensities vs. (a) kinetic energy (logarithmic scale); (b) total energy/rest energy (logarithmic scale); (c) rigidity (linear scale).

or an exponential rigidity spectrum, as can be seen from the figure. The following results were obtained in each case.

Power law of kinetic energy

$$dN/dE=CE^{-\gamma} \qquad \gamma=5.6\pm0.3$$

Power law of total energy

$$dN/dE=C(W/W_0)^{-\gamma} \qquad \gamma=47\pm3$$

Exponential rigidity

$$dN/dP=C\exp(-P/P_0) \qquad P_0=52\pm3$$

In an attempt to distinguish between these spectral forms the integral spectra are plotted in figure 4 together with the low energy points obtained from counters on the Injun I satellite (Maehlum and O'Brien, 1962) the point at 1.5 Mev (53 Mv rigidity) was plotted assuming that the differential intensity between 1.5 and 17 Mev was essentially the same as the integral intensity above 1.5 Mev. Figure 3 shows that an extrap-

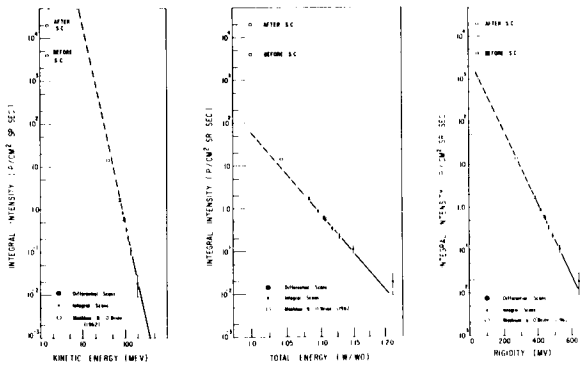


FIGURE 4.—Integral proton intensities versus (a) kinetic energy (logarithmic scale); (b) total energy/rest energy (logarithmic scale); (c) rigidity (linear scale).

olation of the exponential rigidity spectrum, suggested by Freier and Webber (1962) fits the Injun I data most closely. However, the power law of total energy is not inconsistent with the point at 40 Mev, though it does not fit the point at 1.5 Mev. The power law of kinetic energy predicts much higher intensities than are observed at the lower energies.

OBSERVATIONS ON JULY 18, 1961 (FLIGHT 3)

A large stack of nuclear emulsions was flown on July 18, 1961. Originally intended to study the particle intensities very late in the event of July 12, the flight was actually launched at a time which almost coincided with the three plus flare that caused the solar particle event of July 18 and floated at a mean ceiling altitude of 2 gm/cm_2 for a six hour period when the 30 mc/s polar riometer at Churchill was registering its maximum absorption for the event.

The electron sensitive emulsions from this flight were essentially saturated and only an integral proton intensity at 100 Mev of 40 ± 8 p/cm^2 sr sec has been obtained. In the future the less sensitive emulsions which were in the stack will be used for a composition study of the low energy solar particles.

DISCUSSION

The Spectra

The spectra which are obtained from the emulsion data of flight 2 are averaged over a ten hour period and have meaning only if the shapes of the spectra did not change markedly during this time. In this period there was a sudden commencement at 1112 UT on July 13, and the 27.6 mc riometer at College (Leinbach, 1962) showed an increase in absorption during the magnetic storm following the sudden commencement. Counter results from the State University of Iowa satellite Injun I (Maehlum and O'Brien, 1962) indicate that the increase of absorption at College can be explained by an increase in the intensity of particles below energies of 17 Mev combined with a decrease in the threshold rigidity at the latitude of the College riometer station. The integral flux of particles with kinetic energy greater than 40 Mev as measured by Injun I during the time of the balloon flight remained essentially constant, though there was a slight increase (about 20%) after the sudden commencement. One would expect the particle intensities at higher energies to be affected even less. As a consequence, it is reasonable to assume that the intensity of particles with energies greater than 77.5 Mev was not markedly affected by the sudden commencement and magnetic storm, and that the spectra obtained here are meaningful averages for the time period 0839 UT to 1819 UT on July 13.

For the event of July 18, the emulsion data consists only of an integral point at 100 Mev. The Injun I data (Maehlum and O'Brien, 1962) show that, during the time of the flight, 1305 UT to 1918 UT on July 18, the integral intensity of particles above 40 Mev was going through a broad maximum, increasing from 3×10^2 to 5×10^2 p/cm^2 sr sec. Also, the 30 mc/s riometer at Churchill showed a sustained maximum absorption of 7 to 8 db. This would indicate that the particle intensities were not changing rapidly during the flight period and that the integral flux obtained is a reasonable average value.

Comparison with the Events of April 29 and May 4, 1960

As discussed by Maehlum and O'Brien (1962), the spectra for the two events at the time of maximum riometer absorption are significantly different. The event of July 12 has a very steep spectrum extending to energies well below 100 Mev, whereas the event of July 18 has a significantly flatter spectrum, particularly at energies less than 100 Mev.

The two events discussed here show a strong resemblance to the events of April 29 and May 4, 1960. In the event of April 29, 1960 the Explorer VII counters indicated an omnidirectional flux for particles with energy >30 Mev of 18 p/cm^2 sec. at about 0000 UT on 30 April (Lin, 1961). At this time the 27.6 mc/s riometer at Thule was showing absorption of about 8 db (Leinbach, 1962). One must assume, in order to explain the large riometer absorption, that the spectrum for the April 29 event was quite steep and had a large number of particles with energies below 30 Mev. Hence, the shape of the energy spectrum must have resembled that of the July 12 event.

TABLE II.—*Comparison of the Events of July* 1961 *and April, May* 1960

1615 U.T. 11 July	Class 3 flare, S6E32 Small PCA	before 0130 U.T. 28 April	Class 3 flare, S5E34 Small PCA
1000 U.T. 12 July	Class 3+ flare, S7E22 Large, slow developing PCA Steep solar proton spectrum	0138 U.T. 29 April	Class 2+ flare, N12W20 Large, slow developing PCA Steep solar proton spectrum
1112 U.T. 13 July	Sudden commencement Forbush decrease Rapid increase in absorption	0132 U.T. 30 April	Sudden commencement Forbush decrease Rapid increase in absorption
≃0730 U.T. 14 July	Resurgence of magnetic storm	1213 U.T. 30 April	Sudden commencement Forbush decrease
0920 U.T. 18 July	Class 3+ flare, S8W60 Neutron monitor increase Fast rising event Flat solar proton spectrum	before 1000 U.T. 4 May	Class 2 flare, N14W90 Neutron monitor increase Fast rising event Flat solar proton spectrum
0247 U.T. 20 July	Sudden commencement No Forbush decrease	≃1650 U.T. 6 May	Magnetic storm No Forbush decrease

Flare Times are taken from National Bureau of Standards CRPL–F Reports, Part B.
Sudden Commencement times are from J. Virginia Lincoln, Selected Geomagnetic and Solar Data, J. Geophys. Research **65,** 4195, 1960 and **67,** 381, 1962.

During the event of May 4, 1960 there was a neutron monitor increase of over 300% at Deep River (Carmichael and Steljes, 1961) though the 27.6 mc/s riometer at Thule showed a maximum absorption of only 5 db (Leinbach, 1962). In addition the spectrum measured between 1700 UT May 4 and 0200 UT May 5 by Biswas and Freier (1961) was very flat.

For each of these periods there was a flare followed by a small PCA event, a flare about a day later followed by a large PCA event and solar particle event characterized by a steep spectrum, followed in turn a few days later by a flare and solar particle event characterized by a flat spectrum. These and other similarities between the two events are summarized in Table II. This indicates that conditions in the space between the sun and the earth and on the sun were very similar during April and May 1960 and July 1961 in those respects which are important in the production of accelerated particles and their transit to the earth. A more detailed intercomparison may indicate just which are the important conditions.

ACKNOWLEDGMENTS

We are grateful to Dr. C. E. Fichtel for moral and physical encouragement during the course of this work. One of us (CJW) wishes to thank the University of Bristol for leave of absence while holding the NAS-NASA Senior Postdoctoral Research Associateship.

REFERENCES

Biswas, S., and Freier, P. S., The High-Energy Cosmic Ray Flare of May 4, 1960. 2. Emulsion Measurements, *J. Geophys. Research*, **66**, 1023, 1961.

Carmichael, H., Private communication.

———, and Steljes, J. F., Review of Recent High Energy Solar Particle Events Including November 1960, Proc. Kyoto Conference, II, 337, 1961.

Freier, P. S., and Webber, W. R., Exponential Rigidity Spectra for Solar Flare Cosmic Rays, submitted to *J. Geophys. Research.*

Leinbach, H., Interpretation of the Time Variations of Polar Cap Absorption Associated with Solar Cosmic Ray Bombardments, Geophysical Inst., University of Alaska Report UAG–R127, 1962.

———, The Polar Cap Absorption Events of July 11–20, 1961, Geophysical Institute, University of Alaska Report UAG–R126, 1962.

Lin, W. C., Observation of Galactic and Solar Cosmic Rays from October 13, 1959 to February 17, 1961 with Explorer VII, State University of Iowa Report SUI–61–16.

Lincoln, J. Virginia, Selected Geomagnetic and Solar Data, *J. Geophys. Res.*, **66**, 3949, 1961; *J. Geophys. Research*, **67**, 381, 1962.

Maehlum, B., and O'Brien, B. J., Solar Cosmic Rays of July 1961 and their Ionospheric Effects, *J. Geophys. Research*, **67**, 3281, 1962.

National Bureau of Standards, Solar Geophysical Data, CRPL–F Reports, Part B, 1961.

Appendix I

Assuming the incoming protons have a differential energy spectrum of the form

$$dN = -CE^{-\gamma}\, dE$$

then if R is the range of material traversed from the top of the atmosphere $E = aR^n$ where a and n are essentially constants over a reasonable spread of energies

Thus

$$dN = -C(aR^n)^{-\gamma}\, anR^{n-1}\, dR$$

Now at one level in the emulsion the density of particles coming to rest, ΔN, is given by

$$\Delta N = \int_{R1}^{R_{\max}} \sin\theta \exp(-R/\lambda)\, dN$$

where R_1 is the vertical range, θ the zenith angle and λ the absorption mean free path, so

$$\Delta N = -Ca^{(1-\gamma)}n \int_{R1}^{R_{\max}} \sin\theta R^{(\lambda-1-n\gamma)} \exp(-R/\lambda)\, dR$$

but $R = R_1 \sec\theta$, so

$$\Delta N = -Ca^{(1-\gamma)}\, \mathrm{N}R_1^{n(1-\gamma)} \int_0^{\pi/2} (\sec\theta)^{n(1-\gamma)} \tan\theta \times \sin\theta \exp(-R_1 \sec\theta/\lambda)\, d\theta$$

Thus at two different levels with vertical ranges R_1 and R_2, *if* we can neglect the effects of nuclear absorption

$$\Delta N_1/\Delta N_2 = (R_2/R_1)^{n(\gamma-1)}$$

THE WORLD MAGNETIC SURVEY

J. P. HEPPNER

Goddard Space Flight Center

The mathematical and graphical description of the earth's main field has been, and is, a "data limited" problem. The World Magnetic Survey (WMS) is an endeavor to minimize this limitation by rapidly and comprehensively blanketing the earth with magnetic field measurements. Satellite surveys, which will play a key role in the WMS, are the principal topic of this paper. Existing magnetic field descriptions, the expected results from new surveys, and the methods of obtaining these results with the POGO satellite are emphasized. It is anticipated on the basis of extrapolation from Vanguard 3 results and other considerations that a factor of 10 improvement will be obtained. This means that the average errors of 1 to 3 percent now present in field charts and spherical harmonic descriptions should be reduced to 0.1 to 0.3 percent as a result of the survey.

INTRODUCTION

"The World Magnetic Survey" (the WMS) denotes an international effort to obtain data for mathematical descriptions of the earth's magnetic field and its secular variations and for constructing charts of the field at the earth's surface. It is distinguished from previous survey efforts in the scope of the effort which hopefully will lead to a truly world wide coverage. In obtaining this coverage it is particularly important that: (a) emphasis be placed on the magnetically unmapped and poorly mapped regions of the globe, and (b) the survey be conducted within a time interval that is sufficiently short to not require major corrections for secular variations in describing the field at a particular date. The advantages of conducting the survey during years of minimum solar activity relate the WMS to the International Years of the Quiet Sun (the IQSY). The context of the WMS is logically extended to include studies of temporal magnetic variations at least to the degree that these variations influence the survey measurements. The future determination and study of secular variations is obviously closely related to the primary objectives of the WMS.

The principal objective of the WMS, as first proposed at the IUGG meeting in Toronto in 1957, was to obtain more accurate magnetic charts through expansion of existing surveys and creation of new programs for magnetic surveys by sea and air. The advantages of conducting surveys during years of sunspot minimum and the potential use of rockets and satellites in mapping the field were noted at the meeting of the ICSU Special Committee for the I.G.Y. (CSAGI) in Moscow in 1958 (Chapman, 1960). Resolutions, recommendations, and discussions at these and subsequent international meetings (e.g. IUGG, Helsinki 1960; COSPAR, Florence 1961) gave the WMS an international foundation and established guidelines to promote compatibility of measurements from various sources. These have been described in detail in the IAGA–IUGG, Instruction Manual on the World Magnetic Survey, prepared by Vestine (1961).

Since the proposal in Toronto in 1957 the advent of the satellite as a new vehicle for magnetic surveys has been paralleled by new reasons for needing a world magnetic survey. The need in 1957 came primarily from lack of adequate data for studying the origin of the magnetic field and its secular changes, and preparing more accurate magnetic charts for nautical and aeronautical navigation. These reasons are still fundamental but of equal importance now is the need for better descriptions of the earth's magnetic field in space. In space science, this need comes from the

requirements for: (a) an accurate main field reference for satellite magnetic field experiments designed to determine the distribution of sources of magnetic disturbance, (b) describing the motions and distribution of trapped particles and the trajectories of solar particles and cosmic rays as they approach the earth, (c) determining the trapping lifetime for particles naturally or artifically injected into the magnetosphere, and (d) determining the detailed geometry of field lines for field dependent radio transmissions, for locating conjugate points, etc. In space technology, this need arises from the use of the magnetic field in spacecraft orientation control and aspect systems.

A magnetic "chart" or "map", a two dimensional representation of the field, would be of very limited use in space problems. A mathematical description on the other hand is directly useful. Thus, though one speaks of "mapping" the field, it should be understood that the objective is first a three dimensional (latitude, longitude, altitude) mathematical description of the field, and second, a four dimensional description in which secular changes are adequately known to write the coefficients as a function of time. A related question, discussed later, is whether or not world magnetic charts at the earth's surface drawn from mathematical descriptions will be more or less representative than charts constructed by contouring measurements as done in the past.

A world magnetic survey, in a strict sense, has never been made and has never been feasible by land, sea, and air if one accepts a definition that such a survey ideally requires global coverage on a time scale short compared to the time in which secular variations invalidate the first measurements relative to the last. This does not mean that land, sea, and air methods have become obsolete as this would be misleading in view of the complementary nature of surface and space measurements and the advantages of surface data in mapping crustal anomalies. It does however imply that satellites have made a truly world magnetic survey feasible and, if successful, it may in the future be fitting to refer to this survey as "The First World Magnetic Survey". Logically this survey will be followed by others to update the descriptions of the magnetic field and delineate the patterns of secular change.

Appropriate to this journal, satellite magnetic surveys are the principal topic of this review. Land, sea, and air surveys are briefly discussed to give perspective on the surface data available for main field analyses and to illustrate the complementary features of the various surveys. Existing magnetic field analyses are discussed. Attention is then directed to: problems in satellite surveys, details of the magnetic survey to be conducted with the POGO satellite, analysis problems likely to be encountered in arriving at the best description of the earth's field and secular variations from the data obtained, some of the preflight thoughts on methods of handling and analyzing data from POGO, and the many possibilities for studies of temporal magnetic variations using simultaneous data from the survey vehicle and surface observatories. The satellite discussion will center on the POGO (Polar Orbiting Geophysical Observatory) survey largely because details on other satellite surveys are not available. The POGO survey should, however, be adequately illustrative.

No attempt is made to provide a comprehensive list of references. Recent papers are referenced in preference to many outstanding, older publications on geomagnetism on grounds that these can be found by examination of the references in recent publications and Chapman and Bartels' *Geomagnetism* (1940).

LAND, SEA, AND AIR SURVEYS

a. Observatories and Repeat Stations

Measurements on land come primarily from two sources: magnetic observatories and repeat stations. The quantity and distribution of data for the period 1905–1945 is illustrated in the *Carnegie Institute's Publication 578* (Vestine et al., 1947b). As of 1945 there were approximately 100 observatories and 2000 repeat stations where a repeat station is defined as a location occupied more than once. Since 1945 the number of observatories has increased to about 150. A recent listing is available in the *Annals of the IGY*, Vol. VIII, 1959. The number of new repeat stations is not readily estimated accurately but it appears unlikely that the number has greatly increased.

The distribution of observatories and repeat stations is extremely spotty not only as a consequence of the distribution of land masses but also because the locations within land areas tend to be grouped. The fact that the field at a single observatory or station may not be representative of a region, because of a local anomaly, does not permit one to assign large weighting factors to observatory data relative to other observations for mapping purposes. Although these are serious disadvantages, observatories provide the most reliable data for studies of secular change and in the past the only data for studying secular change on a year to year basis. Observatories also provide the data for editing and correcting survey measurements for temporal variations when these corrections are applied.

b. Ship Surveys

The earliest magnetic charts, epoch 1700, were those of declination drawn from ship records of compass direction. Although magnetic measurements were made on various ocean expeditions in the 1800's (Fleming, 1939) it was not until 1905 that a serious attempt was made to map the oceans. The most comprehensive surveys by ship were those made by the *Carnegie*, a non-magnetic ship constructed as a part of the Carnegie Institution of Washington's program for magnetic mapping conducted from 1905 to 1929 when the ship was destroyed by fire. The extensive coverage achieved by the Carnegie Institution's 25 year survey has been illustrated by Fleming (1939) and Vestine (1960). Ship surveys were reactivated in 1956 by the USSR non-magnetic ship, the *Zarya*, which in 1957 and 1958 covered 47,000 nautical miles in the Atlantic and Indian Oceans (Ivanov, 1960, 1961) and has since surveyed also in the Pacific. Using towed magnetometers measuring the scalar field, ship surveys have also been conducted on a regional scale by various groups engaged in oceanographic and submarine geologic studies and magnetic traverses have been made on various expeditions.

The relative merits of ship and aerial surveys are often compared. Such comparisons are often confused by lack of distinction between what has been achieved and what it is possible to achieve. It is obvious that ship surveys take considerably more time and global secular changes can only be studied in terms of a number of decades. There has, however, been a general belief that this disadvantage is partially counterbalanced by greater system accuracy in ship surveys because they are less susceptible to navigational errors. This belief may be justified in given cases but generalizations are not obviously justified.

c. Aerial Surveys

Magnetometers carried or towed by aircraft and measuring the scalar field have been used for local and regional geologic studies by commercial firms and national geological surveys for about 15 years. The data from these efforts have in some cases been available for constructing magnetic charts. About 10 years ago groups in the United States and Canada redesigned the airborne systems to provide vector measurements and later undertook magnetic surveys on a scale which would contribute more significantly to world magnetic charts. The Canadian Dominion Observatory survey subsequently covered Canada in a nearly uniform pattern and provided several traverses across the Pacific and North Atlantic (Serson, 1960). The U.S. Naval Hydrographic Office in 1959 initiated its plane for world-wide coverage of ocean areas under the name, Project Magnet. The coverage achieved by Project Magnet through February 15, 1963 is illustrated by the heavy lines in figure 1. The lighter lines indicate future survey tracks. The shaded area in the Atlantic Ocean was surveyed between 1953 and 1959.

An excellent review of airborne magnetic surveys for world charts, expressing both the capabilities and the limitations, has been written by Serson (1960). A summary is hardly possible here but it will be instructive to note some of the problems encountered in aircraft surveys for later reference in discussing similar satellite problems. Serson (1960) has listed the sources of error in a vector aerial system as follows: (a) errors in measuring the magnetic field at the magnetometer with respect to the direction reference system, (b) errors in the direction reference system, (c) errors due to changes in the magnetic field of the aircraft, (d) errors due to transient changes in the magnetic field, and (e) errors in geographical

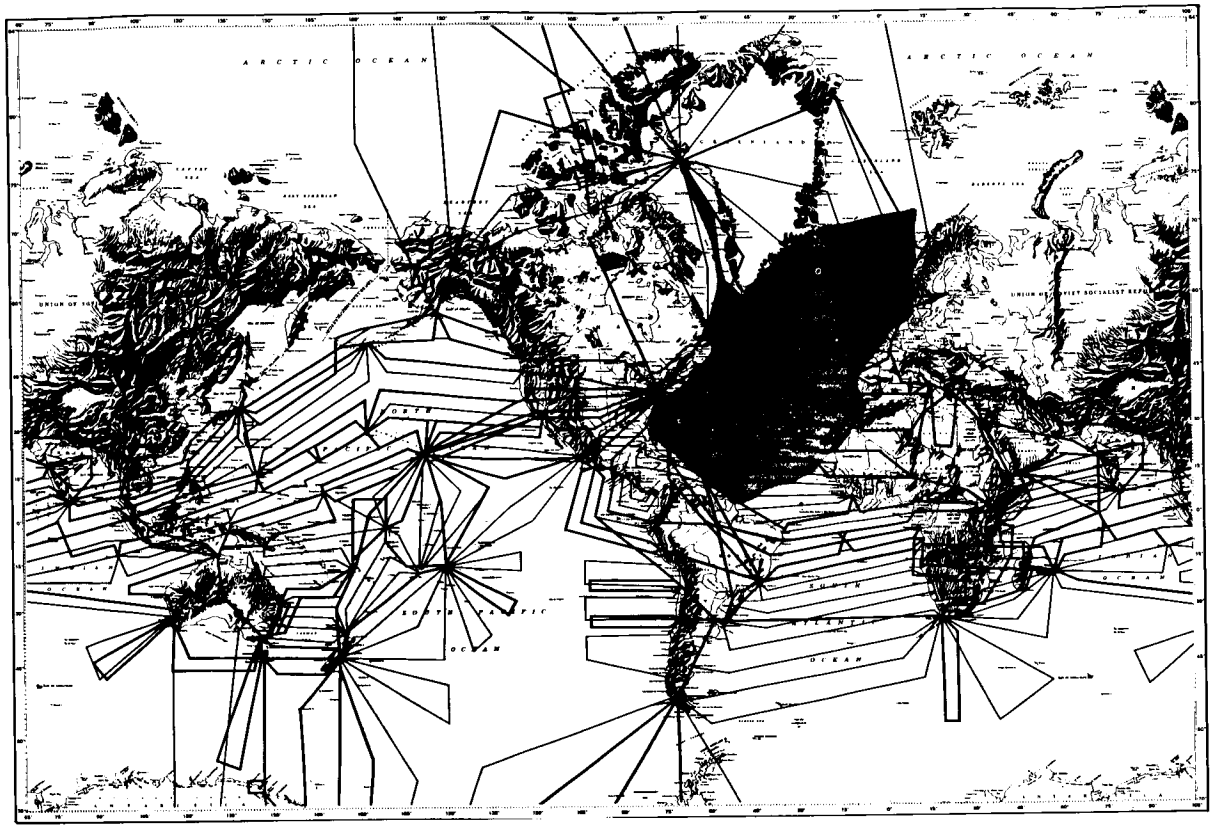


FIGURE 1.—Project Magnet airborne geomagnetic survey tracks: 1959 to February 15, 1963 (heavy lines); future tracks (light lines); surveys between 1953 and 1959 (Atlantic Ocean shaded area). (Courtesy of U.S. Naval Oceanographic Office.)

position. Serson describes the approximate equivalence of the U.S. and Canadian systems. For Project Magnet, Byrnes (1960) gives the following probable errors of observations under good survey conditions: total field magnitude, ±15 gammas; angle accuracies equivalent to inclination and declination, ±0.1 degrees; geographical position, ±5 kilometers. Transformed to component errors the angle errors give north-south and east-west errors varying from zero at the equator to 100 gammas at high latitudes and vertical errors ranging from 50 gammas at the equator to zero at the dippoles. Regarding the geographical position, or navigational, errors Serson (1960) states that errors of the order of 100 to 150 gammas resulting from moderate errors of a few kilometers in navigation are responsible for the largest source of error in the results as plotted on charts. It should be noted, however, that errors of 100 to 150 gammas should not result from position inaccuracies of a few kilometers in the horizontal plane in the absence of crustal anomalies. Thus in a spherical harmonic analysis limited to describing very deep sources the effective field error due to position errors is reduced to values proportional to the regional gradient if the position errors are systematic. If the position errors are random their principal effect is that of increasing the magnitude of the residuals, or noise level, already present in the analysis as a consequence of local anomalies. Returning now to field measurement errors of 50 to 100 gammas these become slightly more important than a few kilometers error in horizontal location for global harmonic analyses. The considerations of systematic *vs.* random errors apply equally well for these errors. Thus, for mathematical descriptions of the field, world charts, and especially studies of secular variations it is desir-

able to reduce the errors in aerial surveys but it is even more important that attention be directed toward avoiding systematic errors.

Another topic worthy of discussion, but not detailed here, is the relative merits of total field *vs.* vector measurements in aircraft surveys. Noting that the errors quoted above are probable values under good survey conditions, the total range of operational errors can be expected to be considerably larger and one can question whether the angle measurements are always of value. The answer at present and at least until methods of analysis of world wide scalar values are thoroughly demonstrated is unquestionably yes. It follows however that there may be considerable merit in flying absolute scalar magnetometers in the same aircraft with the saturable core magnetometers for checking and reducing errors.

EXISTING MAGNETIC FIELD DESCRIPTIONS

a. Spherical Harmonic Analysis of Surface Data

The reader is referred to Chapman and Bartels (1940) for detailed discussion of the applicability of spherical harmonics to field analyses. The basic formulation is given here for later reference. In a source-free spherical shell ($a<r<R$) the magnetic potential, V, satisfies Laplaces equation $\nabla^2 V=0$ and can be expressed in terms of spherical coordinates as follows:

$$V=a\sum_{n=0}^{\infty}\sum_{m=0}^{n}P_n^m(\cos\theta)\times\left[\{c_n^m(r/a)^n+(1-c_n^m)(a/r)^{n+1}\}A_n^m\cos m\phi+\{s_n^m(r/a)^n+(1-s_n^m)(a/r)^{n+1}\}B_n^m\sin m\phi\right] \quad (1)$$

where c_n^m and s_n^m are numbers between 0 and 1 representing the fractional contribution from sources external to R at $r=a$. The $P_n^m(\cos\theta)$ are the partly normalized associated Legendre functions which have been extensively used in geomagnetism since their introduction by Schmidt (see Chapman and Bartels, 1940). As the magnetic force rather than the potential is the measured quantity, harmonic expressions for $\mathbf{F}=-\text{grad}\ V$ are used in analysis, where $\mathbf{F}$ is the total vector field intensity.

Until very recently all analyses were based on the coefficients of one or more of the components

$$Z=\partial V/\partial r \quad \text{(vertical)} \quad (2)$$

$$X=\partial V/r\,d\theta \quad \text{(north)} \quad (3)$$

$$Y=-\partial V/r\sin\theta\,d\phi \quad \text{(east)} \quad (4)$$

and used only surface measurements. Thus, at $r=a$, where "a" is the earth radius, some condensation of the harmonic expressions comes from $r/a=1$ and X, Y, and Z can be expressed as follows:

$$Z_{r=a}=\sum\sum(\alpha_n^m\cos m\phi+\beta_n^m\sin m\phi)\,P_n^m(\cos\theta) \quad (5)$$

$$X_{r=a}=\sum\sum(A_n^m\cos m\phi+B_n^m\sin m\phi)\frac{dP_n^m(\cos\theta)}{d\theta} \quad (6)$$

$$Y_{r=a}=(1/\sin\theta)\sum\sum(mA_n^m\sin m\phi-mB_n^m\cos m\phi)\times P_n^m(\cos\theta) \quad (7)$$

where

$$\alpha_n^m=[nc_n^m-(n+1)(1-c_n^m)]A_n^m \quad (8)$$

$$\beta_n^m=[ns_n^m-(n+1)(1-s_n^m)]B_n^m \quad (9)$$

Most of the approaches for obtaining the coefficients have differed only in numerical technique and weighting factors although there are exceptions noted later. The method used by Vestine, et al. (1947a) is illustrative. In this analysis the functions $X_n^m=dP_n^m(\cos\theta)/n\,d\theta$ and $Y_n^m=WP_n^m(\cos\theta)/n\sin\theta$ tabulated by Schmidt (1935) were used for convenience. World magnetic charts compiled separately for X, Y, and Z were used to obtain from X and Y two determinations of A_n^m and B_n^m and from Z the α_n^m and β_n^m coefficients. Data were taken at intervals of 10 degrees in colatitude and longitude. Fourier coefficients a_n^m, b_n^m were first determined for $m+6$ along each 10° parallel of colatitude and then fitted by the functions X_n^m and Y^m to obtain two values for A_n^m and B_n^m (except for zonal harmonics, $m=0$, given only by X). The Fourier coefficients a_m,b_m were weighted as a function of

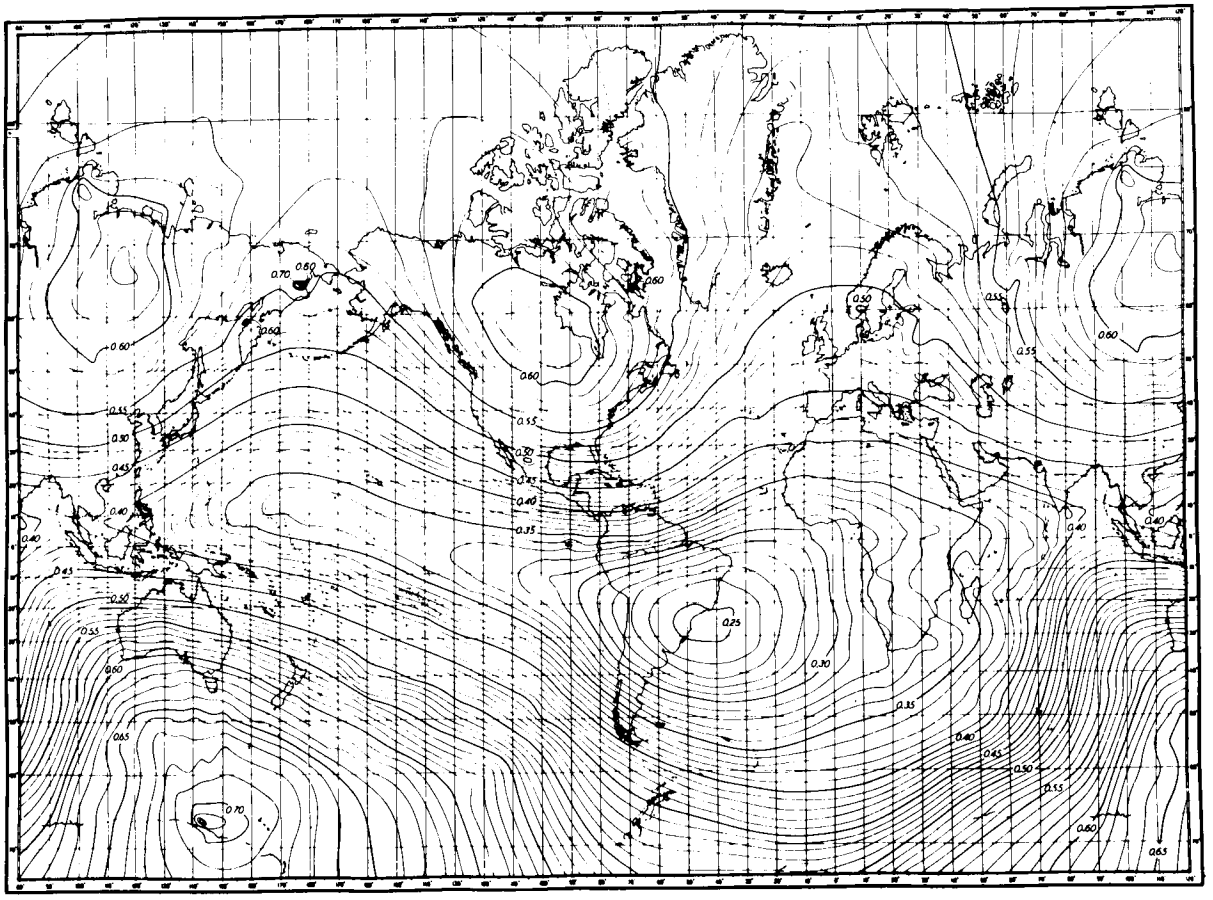


FIGURE 2.—Chart of total magnetic field intensity for Epoch 1945 in units of gauss (Vestine et al., 1947b).

lattitude from 10 for $\theta=80°$, 90°, and 100° to 1 for $\theta=10°$ and 170°. To separate internal and external sources the $\alpha_n{}^m,\beta_n{}^m$ from Z and averages for $A_n{}^m,B_n{}^m$ from X and Y were used in Equations (8) and (9) to give $c_n{}^m$ and $s_n{}^m$.

The above illustrates a number of features which have individually been common to a number of analyses, such as: the use of world magnetic charts for input data, the dependence on Z in separating internal and external sources, the weighting of data as a function of latitude, and the assumption that $r=$constant. For exceptions to several of these feature and detailed differences between analyses the reader should examine the analyses reviewed by Chapman and Bartels (1940) and the more recent analyses of Finch and Leaton (1957), Fanselau and Kautzleben (1958), and Jensen and Whitaker (1960).

In using readings scaled from the world magnetic charts the accuracy of the analysis is necessarily limited to the accuracy of the chart. This, in turn, is a function of the latitude-longitude distribution of observations, the year of the observations relative to the chart year, the measurement errors including temporal variations, and the subjective nature by contouring by hand. The latter point is not at all negligible as illustrated by comparing magnetic charts drawn by different organizations using essentially the same data. Thus, in the past it has been argued that a better description might be obtained using only observatory values. This represents an extreme view which ignores the influence of local anomalies at observatory sites and the fact that the mutual dependence of the coefficients is such that large errors are likely to occur in regions lacking an observatory. Figure 2, by comparison with figure 3, illustrates the detail that appears on

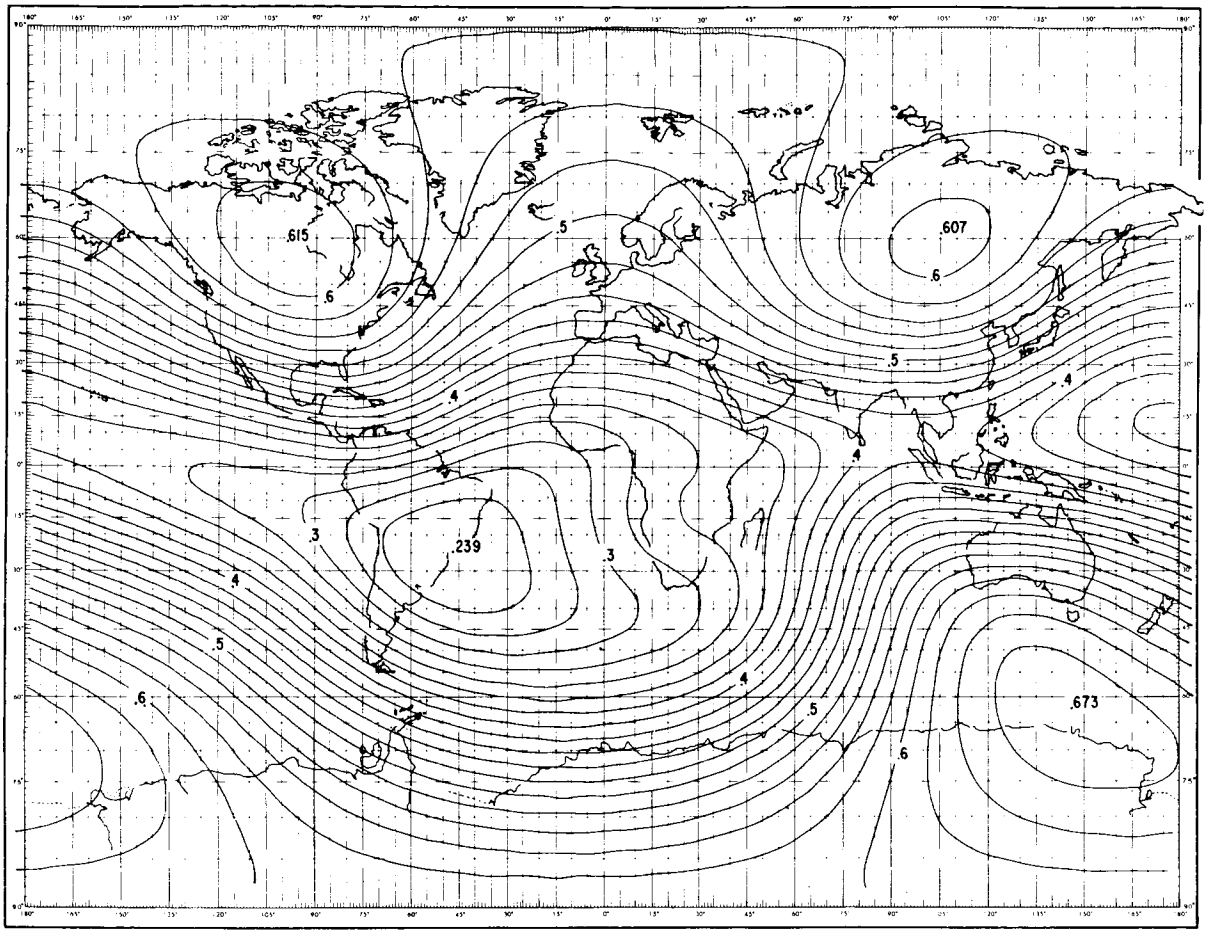


FIGURE 3.—Machine plot (Cain and Neilon, 1963) of total magnetic field intensity for 1960 given by 48 coefficient expansion of Jensen and Cain (1962).

a world chart, after corrections and analysis, relative to the detail given by an $m=n=6$, 48 coefficient, representation of the field. When expanded to greater dimensions and broken into smaller contour intervals the detail in figure 2 increases (see Section maps of Vestine et al., 1947b) whereas the detail in figure 3 remains constant. If the small wiggles and bumps in the contours of figure 2 are the consequence of local anomalies not being perfectly removed from the observations (which is very likely: see discussion in Section 3c; also note the prevalence of irregularities in poorly mapped regions) they will create errors in any spherical harmonic analysis based on chart scalings at regular grid intervals. The use of regular grid intervals is also subject to criticism on grounds that poorly mapped and unmapped regions receive the same weight as regions that are well mapped. The availability of high speed digital computers is currently providing a compromise between the various extremes. An analysis program which utilizes directly all the data that is normally used in chart making can eliminate the subjectivity of the hand contouring without being influenced by local anomalies and data distribution to any greater degree than in the contouring process. Initial steps for achieving this compromise have been made in the methods of Jensen and Cain (1962) and Fougere (1963).

Efforts to determine c_n^m and s_n^m of equations (8) and (9) have placed an upper limit of several percent on the fraction of the surface field caused by external sources. Various errors may contribute to give this large upper limit but it is most

probable that the accuracy is primarily limited by average errors of 1 to 3 percent in the charts used for input data. Another factor is that charts for Z are in general likely to be the least accurate of the three components. The sensitivity to the accuracy of the Z charts is illustrated by an example in Chapman and Bartels (1940, p. 664) in which a 0.3 mm displacement of isoclines of inclination (I) is shown to be equivalent to a 3 percent external contribution found in an analysis by Bauer (1923). In analyzing the vertical component of both U.S. and U.S.S.R. charts for 1955 Jensen and Whitaker (1960) noted differences of up to 6 percent in some regions. Ivanov (1961) in reporting on *Zarya* measurements noted regions where the errors in the U.S. charts for Z exceeded 2000 gammas, approximately 8 percent. Ivanov (1961) also states that the discrepancies between U.S. charts for the total scalar field, F, and the *Zarya* measurements amount to 1500 gammas in almost the entire Southern Hemisphere at latitudes less than 40 degrees and speculates that the errors are probably greater at higher latitudes. The *Zarya* in general showed better agreement with the British Admiralty charts in the Atlantic and Indian Oceans at latitudes less than 40 degrees. Errors in the British charts approached 1 degree in declination and 500 gammas in intensity in the Northern Hemisphere and 2 degrees and 1000 gammas in the Southern Hemisphere.

The data limitations also explain why the assumption of a constant radius, $r=a$, which neglects Δr's of 1 part in 297 from earth oblateness as well as topographical differences, has had little, if any, influence on results.

b. Recent Analyses Using *F* and Space Data

The need for methods of analysis utilizing total scalar magnetic intensity, (F), rather than component intensities appeared about 1957 when it became apparent that data from future satellites would be primarily scalar as a consequence of the instruments to be used and even more important the difficulties of achieving the accurate reference directions required for component measurements. Zmuda (1958) first illustrated that the Gaussian coefficients g_n^m, h_n^m for internal sources appearing in

$$V=a\sum_{n=0}^{\infty}\sum_{m=0}^{n}(a/r)^{n+1}P_n^m(\cos\theta)\times[g_n^m\cos m\phi+h_n^m\sin m\phi] \quad (10)$$

were obtainable through analysis of F^2 by squaring the series for each component X, Y, and Z and adding the terms, thus obtained a series for

$$F^2=X^2+Y^2+Z^2 \quad (11)$$

As illustrated for an eccentric dipole model the first eight harmonics, through $n=2$, $m=2$, give 31 new coefficients representing combinations of the eight desired. Although 31 coefficients is a considerable reduction from the 93 obtained without addition it is apparent that the series becomes immense in higher degree representations such as $n=6$. Zmuda (1958) further treated the problem of separating internal and external sources, but here made an error in assuming that a scalar subtraction of fields due to internal and external sources was a valid simplification.

Jensen and Cain (1962) using expressions (10) and (11) and considering internal sources only, such that the coefficients of Equations (5), (6), and (7) become the Gaussian coefficients g_n^m, h_n^m, devised a method to determine the coefficients which avoids the nonlinear dependence of F^2 (or F) on g_n^m, h_n^m and permits the use of least squares fitting to the data. The method is one of improving on an existing set of coefficients without being dependent on the existing set being very accurate. The procedure is to expand F^2 into a Taylor series that includes only the linear term, giving

$$F^2=F^2{}_0+2\sum[u_n^m(\Delta g_n^m)+v_n^m(\Delta h_n^m)] \quad (12)$$

where

$$u^m=(a/r)^{n+2}\{Z_0(n+1)(\cos m\phi)P_n^m(\cos\theta)-X_0\cos m\phi[dP_n^m(\cos\theta)/d\theta]+Y_0(m\sin m\phi)[P_n^m(\cos\theta)/\sin\theta]\} \quad (13)$$

and

$$v_n^m=(a/r)^{n+2}\{Z_0(n+1)(\sin m\phi)P_n^m(\cos\theta)-X_0\sin m\phi[dP_n^m(\cos\theta)/d\theta]-Y_0(m\cos m\phi)[P_n^m(\cos\theta)/\sin\theta]\}. \quad (14)$$

Values for F_0, Z_0, X_0, and Y_0 are computed at the data locations using the best available set of coefficients g_n^m and h_n^m. The Δg_n^m and Δh_n^m are the corrections made to the initial g_n^m and h_n^m to obtain the best fit to the measurements. The summation term in Equation (12) given by

$$E=F^2-F_0^2 \tag{15}$$

indicates the error, or residual between measured and computed values, and as this term is linear with respect to the Δg_n^m, Δh_n^m a least squares fitting can be used to find the corrections and thus generate an improved set of coefficients. In application the process is one of successively applying the corrections and repeating the procedure until significant improvements in accuracy are no longer obtained. The procedure also applies to the components and can be used to mix component and total field data (e.g., for the horizontal intensity H one has $E=H^2-H_0^2$).

The method of minimizing $F^2-F_0^2$ was first applied to approximately 3000 measurements of the scalar field taken by Vanguard 3 between latitudes±33.5 degrees and altitudes 510 to 3753 km. Although this produces a good fit automatically weights the data using least squares by the factor $(F+F_0)^2$ as seen by rewriting the squares as

$$\sum(F^2-F_0^2)^2=\sum(F-F_0)^2(F+F_0)^2 \tag{16}$$

which emphasizes the data at low altitudes and high latitudes relative to the high altitude and low latitude data. As the difference between $2F$ and $F+F_0$ is negligible in a weighting sense the weighting factor can be removed by minimizing the error quantity

$$E=(F^2-F_0^2)/2F\approx F-F_0 \tag{17}$$

in place of (15). Another alternative is to minimize the fractional (or percentage) error given by

$$E=(F-F_0)/F. \tag{18}$$

This has the effect of weighting the high altitude or low latitude data more heavily than the low altitude or high latitude data.

In subsequent analyses of Vanguard 3 data (Cain et al., 1962b) all three expressions (15), (17), and (18) were used with comparable results. Average residuals *vs.* altitudes from use of (15) and (18) are shown in figure 4. The series involved 63 coefficients, g_n^m and h_n^m with n and m of 7, which, as an example taking expression (18) fitted all data (disturbed as well as quiet days) with a root mean square residual of 21 gammas or 0.13 percent. Of this 12 gammas could be attributed to errors of measurement coming primarily from orbit position errors. The remaining 9 gammas was easily attributed to disturbance effects. In contrast use of the initial Finch and Leaton (1957) coefficients on the same data gave an r.m.s. residual of 255 gammas. Thus the method worked well for this limited case which represents strictly a curve fitting procedure not taking into account the existence of external sources and not attempting to fit data outside the regions of measurement. The coefficients for this special analysis of limited regions over North and South America, the Caribbean, Australia, and South Africa are given by Cain et al. (1962a).

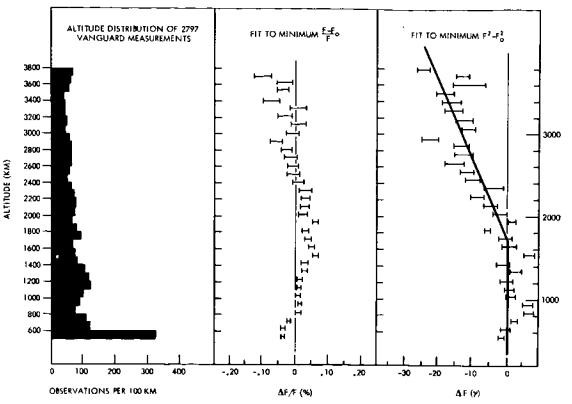


FIGURE 4.—Average residuals of Vanguard 3 data relative to two different, $m=n=7$, spherical harmonic reference fields as a function of altitude. Error bars show the standard errors of the mean residuals.

The method was next applied by Jensen and Cain (1962), using expression (17) and the H equivalent, to a collection of 74,000 values of F and H taken since 1940 of which about 28,000 points were F measurements from aerial surveys, 3000 points were F measurements from Vanguard 3, and the remainder were surface values for H. To shorten the time for each iteration only every tenth point was used in the computation of coefficients. The fit was made to the year 1960 with the assumption that secular variations could be approximated by a linear time dependence.

This was included by substituting coefficients $\Delta g_n^m+(t-1960)\Delta(dg_n^m/dt)$ and $\Delta h_n^m+(t-1960)\times\Delta(dh_n^m/dt)$ for Δg_n^m and Δh_n^m in equation (12). The accuracy achieved as indicated by the residuals was closely related to the distribution of data, giving root mean square errors of: 135 gammas in regions of good coverage, about 600 gammas in regions of poor coverage, and 304 gammas for the total data. As a check on the method, computed values were compared with U.S. chart values along parallels of latitude and in general closer agreement was achieved by using the new coefficients than by using the Finch and Leaton (1957) coefficients which were used for the initial g_n^m, h_n^m. Although the Jensen and Cain method worked well, major improvements over past analyses could not be expected using the existing data.

Comparing also the relative magnitude of the residuals in fitting just Vanguard data and fitting the more heteorogeneous set of surface data, one gets some indication of the refinements in analysis than can be expected from future satellite data relative to handling data from the earth's surface where local anomalies have a major influence.

c. Questions Regarding Field Sources and the number of Coefficients

It is commonly believed that sources of the field interior to the earth's surface are confined to two regions: a primary region located in the earth's core (depth >2900 km), and a secondary region, the earth's crust. There is also considerable evidence (Serson and Hannaford, 1957; Alldredge and Van Voorhis, 1961; Alldredge et al., 1963) that the sources in the crust most typically have linear dimensions of less than 50 kilometers and the number of occurrences of anomalies with greater dimensions decreases rapidly with increasing dimensions up to several hundred kilometers. Anomalies with dimensions greater than several hundred kilometers appear to be very rare. Thus any attempt to represent typical crustal sources in a global spherical harmonic description would be ridiculously cumbersome and these "local" anomalies are more properly treated as noise. It should, however, be noted that strike-slip fault displacements of magnetic anomaly patterns of 1420 kilometers have only recently been found by Vacquiet, et al. (1961) and it is not inconceivable that major geologic features still unknown may contribute to future descriptions. Similarly the available data does not completely rule out the possibility that crustal differences between oceans and continents may have an influence in future analyses. If we ignore these latter possibilities for the time being, the question then is "how many coefficients are required to describe the interior field as observed at and above the earth's surface, exclusive of crustal anomalies?". The question has direct physical significance in understanding the hydromagnetic dynamics of the core. Although a unique determination of interior sources is not possible with surface measurements, it is partly on the basis that coefficients of higher degree and order than $n=m=6$ have historically been unnecessary that the belief has evolved that sources of continental dimensions and sources in the mantle are essentially non-existent. There are of course physical reasons as well for assuming an absence of primary in the mantle (see, e.g. Elsasser, 1950).

Strictly from an analysis standpoint there are grounds for questioning the ultimate adequacy of an $n=m=6$, or 48 coefficient, representation. Jensen and Whitaker (1960) for example found that higher order terms were not insignificant in their analysis and permitted the computer to run through $n=24$, $m=17$, producing 512 coefficients at which point one percent contributions were still being obtained. They concluded quite logically that the large number of coefficients merely reflected the chart errors as the analysis technique assumed a perfect chart. Similarly, Fanselau and Kautzleben (1958) carried an analysis through $n=m=15$ and reached the conclusion that beyond $n=m=6$ the data was too inaccurate to obtain an improvement in the description. This point is simply seen in a rough way by noting the agreements and disagreements between the signs of the three sets of coefficients in Table 1 which were derived from three different data sources. Through $n=4$ all signs agree; for $n=5$, three sign disagreements are noted in the 11 coefficients; for $n=6$, seven sign disagreements appear in the 13 coefficients. It appears logical to conclude that the $n=m=6$ cut-off point in past global analysis is

TABLE 1.—*Values of g_n^m, h_n^m to $n=m=6$ from several analyses*

n	m	Jensen and Whitaker Year 1955*		Finch and Leaton Year 1955		Jensen and Cain Year 1960	
		g	h	g	h	g	h
1	0	30396		30550		30411	
1	1	1955	−5677	2270	−5900	2147	−5799
2	0	2331		2280		2404	
2	1	−5007	3133	−5250	3290	−5125	3312
2	2	−1444	−398	−1370	−210	−1338	−158
3	0	−2946		−2950		−3152	
3	1	5672	1346	5860	1380	6213	1487
3	2	−2181	−407	−2440	−570	−2490	−408
3	3	−788	37	−720	80	−650	21
4	0	−3849		−4150		−4179	
4	1	−4292	−427	−4400	−820	−4530	−1183
4	2	−2307	876	−2270	1210	−2180	1001
4	3	770	296	800	80	701	43
4	4	−269	144	−230	120	−204	139
5	0	1402		2090		1626	
5	1	−3925	60	−3220	−170	−3441	−80
5	2	−990	−688	−1550	−750	−1945	−200
5	3	465	−82	170	230	−61	460
5	4	237	195	340	310	278	242
5	5	19	−39	50	−60	70	−122
6	0	−55		−1500		−1952	
6	1	−1515	695	−930	440	−485	−576
6	2	−1336	−1972	−250	−1660	321	−874
6	3	1833	346	2410	−40	2141	−341
6	4	208	160	160	60	105	−12
6	5	−34	−58	−10	60	23	−112
6	6	37	31	70	10	112	−33

*Continues to $n=24$, $m=17$.

imposed by the data and is not necessarily determined by the real form of the field.

Thus for clues to the answer regarding the number of significant coefficients, prior to analysis of WMS data, we have to look at analyses of measurements that are restricted in distribution but are at least homogeneous in quality and taken within a time span that does not involve large uncertainties caused by secular variations. Alldredge et al. (1963) have attempted to do this by piecing together flight tracks from Project Magnet to simulate a single around-the-world, roughly great circle, magnetic profile. Using F rather than components for reasons of accuracy and completeness, harmonic coefficients in a Fourier series were computed up to the 2000th order. The amplitude of the coefficients dropped off rapidly up to order 7. Beyond the 6th order, the envelope of the coefficients slowly decreased in amplitude to a value of about one gamma for the 2000th harmonic. The break above the 6th order coefficient was attributed to the separation of core and crustal sources. The authors noted that roughly 4 million spherical harmonic coefficients would be required to similarly represent crustal anomalies over the entire earth and proceeded to estimate the number of spherical harmonic coefficients required to represent core sources under the assumptions that the single profile was representative of all great circle profiles and smoothing over 400 km distances would effectively remove crustal sources. Determining the number of

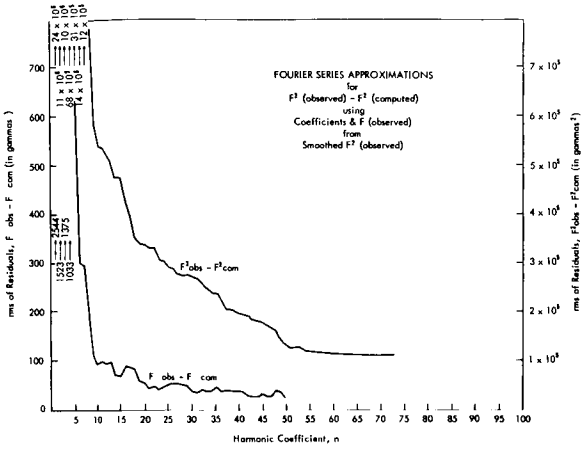


FIGURE 5.—Analysis of a composite Project Magnet around-the-world field track by Alldredge et al. (1963). RMS residuals of F as a function of order of harmonic coefficients for a Fourier series field fit to F^2.

coefficients required to represent F^2, as opposed to F in Fourier analysis, was used as a basis for relating the Fourier and spherical approximations and an intermediate analysis was performed to reduce truncation errors. For these assumptions the authors concluded that spherical harmonic coefficients of $n=m>10$ are probably not required to describe the earth's main field caused by interior sources. The break near $n=10$ is illustrated in figure 5.

Vanguard 3 provides another homogeneous, but restricted in distribution, set of data for approaching the above question. Heppner et al. (1961) illustrated the systematic shift of curves for $\Delta F = F(\text{measured}) - F(\text{computed})$ over several degrees of latitude in the region of Australia at altitudes >510 km (perigee). At perigee, the changes in ΔF approached 0.4 and 0.3 percent, respectively, for Jensen and Whitaker (1960) and Finch and Leaton (1957) coefficients between approximately 25 and 33 degrees south latitude. In this region the percentage change in ΔF with latitude also decreased rapidly with altitude. This behavior was interpreted in terms of internal sources with a rough estimate that harmonics of degree 20 would be required to match the data. The $n=m=7$ analyses of just Vanguard data (see Section 3b) reduced the shift noted above to 0.1 percent but these analyses necessarily put great emphasis on the specific area in question relative to the rest of the world. The possibility exists that the rapid change in ΔF with latitude and altitude could have been caused by a diurnal effect (perigee data was available for night hours only) but the magnitude is incompatible with diurnal variations observed at the earth's surface.

Examination of Vanguard 3 data for much smaller effects than that noted above also leads one to conclude that it is potentially possible to describe the field to an accuracy greater than that permitted by a 63 coefficient, $n=m=7$, representation. Figure 4, for example, shows systematic behavior with altitude despite the smallness of the residuals and it was not found possible (Cain et al., 1962) to eliminate the altitude variation using internal sources only. The negative slope above 1800 km (Figure 4) was present in all data fittings as was the negative displacement in the 500 to 800 km range. Although the distribution of measurements with altitude and local time did not permit Cain et al., (1962) to definitely conclude that this was not a diurnal effect, interpretation in terms of existing theories of diurnal variations and their surface behavior was not promising. A simple alternative, at least until external sources are better understood, is to: (a) attribute most of the negative effect above 1800 km to the influence of external sources in and above the region of measurement, and (b) attribute the negative shift below 800 km to inadequate description of the internal sources with an $n=m=7$ approximation. In detail the negative behavior between 500 and 800 km comes primarily from effects near-perigee in the regions of southern Australia noted above, and an area centered on the southwest corner of the U.S. If explanation (b), above, is correct it suggests that the estimate made by Alldredge et al. (1963), that $n=m=10$ should be sufficient for internal sources, may be too low.

WMS SATELLITES: GENERAL CONSIDERATIONS AND THE POGO SYSTEM

a. Orbit Selection

Orbits will be discussed here in terms of optimum orbits for meeting the basic WMS objectives. Thus, highly eccentric orbits and/or circular at great distances which are primarily intended for

magnetospheric studies will be ignored because they provide relatively little information on the detailed form of the field from interior sources. It should however be recognized that they have considerable importance in determining the distribution of external field sources. Similarly, considering coverage rather than altitude, orbits of low inclination are certainly not in the optimum category for the WMS.

The desirability of having a highly inclined orbit plane, preferably passing close to the poles, is obvious. The need for latitude coverage comes not only from considering the analysis of satellite data but also from the consideration that surface data is very sparse and difficult to obtain in the high latitude regions—especially in the Southern Hemisphere.

If all magnetic field sources were interior to the earth's surface the orbit having the lowest possible altitude would be optimum for the WMS. "Lowest possible", however, still means an altitude that would provide adequate lifetime for achieving the desired global coverage. It also means an altitude at which variations in satellite drag are sufficiently small to not introduce significant errors in the determination of the instantaneous location of the satellite at the time of each field measurement.

The question "what is an adequate lifetime and/or coverage?" has two quite different answers with the choice depending on the importance attached to removing possible effects of diurnal magnetic variations. The first answer either neglects diurnal effects and hopes that they are negligible or assumes that they can be removed in analysis using independent surface observations in conjunction with a theoretical model for their cause. For this answer, for a polar orbit, one is concerned with the choice of the local time zone of the orbit plane but treats the lifetime solely in terms of the rate of measurement and longitude differences in equatorial crossings to arrive at a measurement grid with some minimum spacing between points. The IUGG manual, Vestine (1961), for example, suggests a minimum density of readings each 100 km along the orbit and equatorial crossings separated by a maximum of 400 km. Assuming that the orbit period does not have an integral multiple giving 24 hours, which would give equatorial crossings at the same longitudes on successive days, a grid of this density is easily achieved in several weeks. The exact lifetime depends of course on the orbit period and what portion of the data is considered questionable for mapping because of magnetic disturbance. The second answer comes closer to perfection by assuming that the diurnal magnetic variations are not negligible and similarly that corrections based on theoretical models for their their cause may be grossly erroneous. In this case satellite passes through all local time zones are needed and this requires a lifetime of six months.

Next, to be optimum the orbit should have characteristics which simplify the problem of separating internal and external sources. Here one is concerned, relative to existing knowledge, with external sources having world wide effects such as those of magnetospheric compression by the solar wind and trapped particle drifts and diamagnetism. There is considerable evidence that these external sources contribute at least 20 to 50 gammas, and possible more, to the average surface field. In global descriptions the external field effects have obviously been hidden in the limitations of past analysis (Section 3) but they will assume importance in future analysis anticipating a factor of 10 improvement. It is apparent that the resolution in separating the terms

$$(a/r)^{n+1} \quad \text{and} \quad (r/a)^n$$

in Equation (1) is improved considerably when satellite measurements are available at different altitudes. Thus, Kalinin (1961) suggested that two satellites in circular orbits of different altitudes should be used for the WMS. As noted in the POGO discussion below, a range of altitudes can also be obtained at all latitudes using only one polar satellite provided the orbit is eccentric. One is then faced with the question of an optimum apogee altitude, assuming that a choice of low perigee has been made in accord with the previous discussion. Ideally to get maximum resolution in separating sources one wants the apogee altitude below but approaching close to the altitude of the lowest altitude external source of significance. Conversely one does not want apogee so high that sensitivity to the high order internal terms is lost, particularly if the operating lifetime is shortened by failures. Obviously compromises and assump-

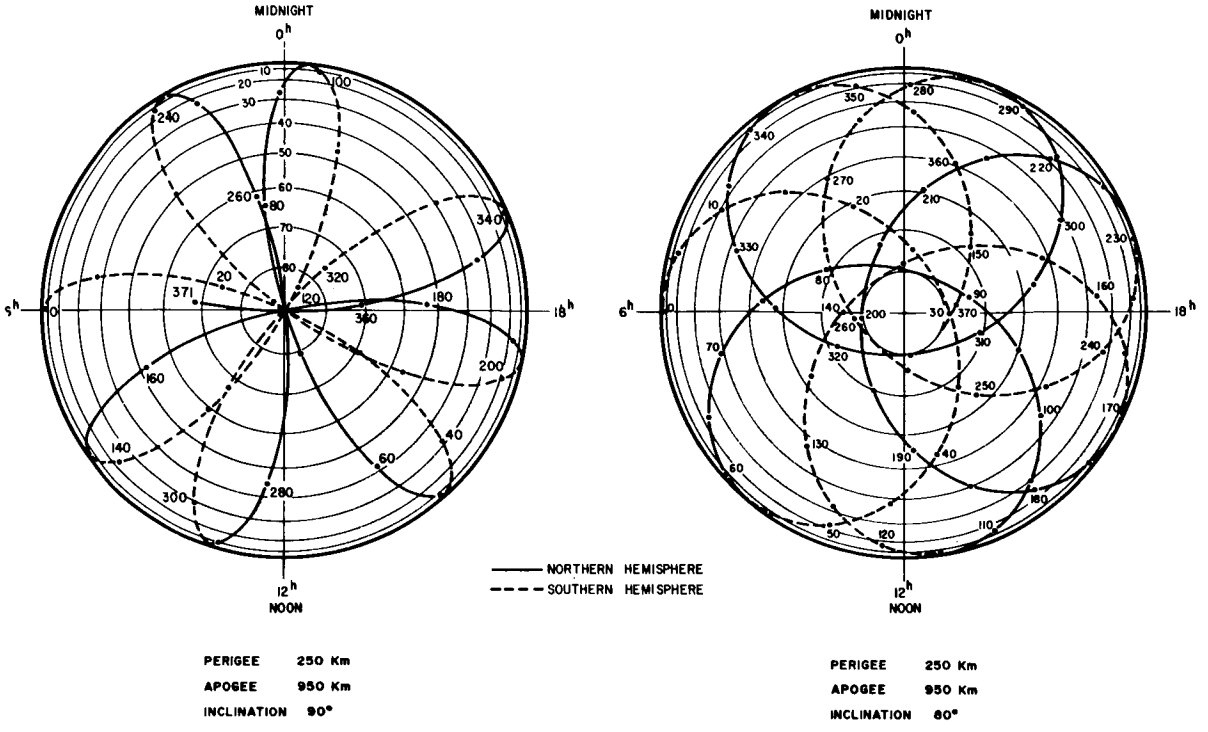


FIGURE 6.—Latitude and local (solar) time of perigee point as a function of days after satellite injection for two orbits similar to the POGO orbit. Local time of injection is arbitrarily taken at 06 hours. Days following injection are noted at 10- or 20-day intervals.

tions become necessary in picking apogee. For example, from the discussion of Vanguard data in Section 3c one might assume a maximum apogee of less than 1800 km for staying below significant external sources and less than 800 km for maintaining sensitivity to interior sources of high order. Similarly one might assume that there was merit in staying below the intense regions of the inner radiation belt (*i.e.* < 1200 km). There is not an obvious select altitude but there are good reasons for keeping apogee in the range of several hundred kilometers above perigee to altitudes less than 2000 km and additional arguments for placing apogee below 800 km if the operating life is short compared to the time it takes for perigee and apogee to move roughly 180 degrees in latitude.

Without going into detail it is apparent that analogous problems are encountered in Kalinin's (1961) suggestion of two circular orbits; for example, in picking the altitude of the highest orbit and in considering the consequences of an early failure in either of the two satellites.

In summary the above depicts the most favorable single orbit as an orbit with the following characteristics: polar inclination, a minimum lifetime of 6 months, perigee as low as possible subject to the lifetime requirement, and an apogee altitude, less critically defined, but located at least several hundred kilometers above perigee and less than 2000 kilometers altitude. There are various other orbits which could give more definitive results pertaining to isolated WMS objectives but these could not individually meet the overall objectives as well as the orbit described above. Further improvements can of course be obtained using two or more satellites in orbits selected to complement each other.

The POGO orbit parameters, as currently planned, are: inclination between 82 and 90 degrees, 250 km perigee, 925 km apogee, and an active life that will approach one year if there are not critical failures. The fact that these characteristics are favorably matched to the optimum single orbit described previously is partly circumstantial in that the initial selection of altitudes for

POGO was primarily the result of other considerations such as a favorable altitude range for a variety of aeronomy experiments.

Figure 6 illustrates the retrograde motion of perigee in days as a function of latitude and local time for two orbits, which are roughly similar to orbits planned for POGO. The assumption of satellite injection, or 0 day, at 06 hour is arbitrary. The equivalence to having a low altitude circular orbit, and a number of higher circular orbits, is evident by noting that within 49 and 58 days, respectively for 90 and 80 degree inclinations, perigee has moved over the entire range of north and south latitudes. An additional feature of a noncircular orbit, pertinent to the external source problem, is that measurements are obtained at two altitudes each day with the exceptions of polar regions and times when the apsides are at maximum inclination.

b. Location Accuracy

Knowing the exact space location (latitude, longitude, and altitude) at the time of each field measurement is one of the most critical problems for the WMS. The instantaneous location of the satellite must be known with errors <1.0 km horizontally and <0.25 km vertically to be consistent with the objective of maintaining maximum field errors from all sources below 10 gammas. This assumes that all other sources of error are small relative to the errors created by the displacement between the true location and the computed location.

Location errors have, in general, been difficult to estimate in an absolute sense because the estimates are usually based on relative agreement between independent determinations. Confidence is developed primarily by finding agreement between determinations employing different types of tracking data. When the satellite is also making absolute magnetic field measurements there is an additional check on the daily, weekly, or monthly precision of the orbit analysis as illustrated by Heppner *et al.* (1960a, 1960b) for Vanguard 3. The critical evaluations of the Vanguard 3 orbit determinations (see Cain et al., 1962a, 1962b) illustrate the problem as they are based on both optical tracking and Minitrack, radio interferometer, tracking and have the additional feature that they were checked against field measurements and subsequently re-evaluated a number of times to remove obvious discrepancies. The results show that the WMS objective would not be met using tracking and analysis techniques similar to those of Vanguard 3.

An important step toward solving this problem in POGO is the inclusion of a two-way doppler, range and range rate, tracking system to supplement the standard radio interferometer, Minitrack, tracking system. This combination of distance, velocity, and angle measurements is expected to provide the required accuracy. Early flight tests of the range and range rate system have for example shown precisions equivalent to maximum errors of several tens of meters. There are, however, other considerations in addition to tracking errors, such as: (a) relative time errors between stations, (b) errors in knowing the exact geodetic location of each station, (c) errors which will vary with satellite location and/or time coming from inaccuracies in knowing or accounting for small differences in the gravitational acceleration, satellite drag, radiation pressure, and in some systems, the propulsion effects of satellite orientation controls, and (d) errors in the methods of analyzing tracking data and computing instantaneous positions over the entire earth when tracking data is available from only a small number of stations. Cumulatively, these considerations cast doubt on any anticipation of maintaining maximum errors well below 10 gammas in a low altitude orbit. Thus, it is likely that location errors will be the limiting factor in the accuracy of the WMS satellite data.

c. Scalar or Vector Measurement Question

A question which has frequently appeared in discussions of the WMS is whether or not the satellite scalar total intensity measurements will be adequate for obtaining a vector description of the field. As the potential derived from scalar measurements in a source free region is unique (see e.g., proof by Jensen, 1962) the question is primarily directed toward the computational feasibility. What little experience there is to date does not suggest that the scalar analysis involves inherent difficulties. In fact, the analyses of Jensen and Cain (Section 3b) suggest that the

question of scalar or vector measurements is not particularly critical. This may, however, be partially a consequence, and perhaps a merit, of the method in that an initial field based on vector information is used as a starting point. Conversely an initial field based on a previous scalar analysis could also serve as a starting point so this can hardly be considered an essential feature at present. It should, however, be recognized that the fields derived from scalar values have to date only considered internal sources and adequate attempts to compare analyses under conditions of external field sources have not been made. Such comparisons would also be relatively meaningless using existing data.

The question is also equivalent to asking whether or not the coefficients derived from the scalar field, F, will give convergence to a higher, equal, or lower noise level, or magnitude of residuals, then a set of coefficients derived from X, Y, Z, or equivalent angle and component measurements. Thus the question in a practical sense is related to the relative accuracy of scalar and component measurements from a low altitude satellite.

Absolute accuracies of a fraction of a gamma to several gammas, depending on the particular system, are readily achieved with nuclear and atomic resonance magnetometers measuring the scalar intensity. It is more difficult to estimate an accuracy for component measurements because a number of variables have to be considered, such as: (a) the accuracy of the magnetometer measuring components or angles, (b) the accuracy of knowing the orientation of a coordinate system at the magnetometer relative to the coordinates of the satellite system providing measurements of the satellite orientation, and (c) the accuracy of measuring the satellites orientation relative to earth coordinates.

A proper review of item (a), which would consider various possible magnetometer systems many of which are in turn dependent on the type of satellite motion, would be unduly long. The use of fluxgate, or second harmonic saturable core, magnetometers is well known: for component airborne measurements (Serson, 1960; Schonstedt and Irons, 1955), for scalar field rocket measurements (Maple et al., 1950; Singer et al., 1951), for scalar field satellite measurements (Dolginov et al., 1960, 1961, 1962). Although item (c), above, would not permit useful vector measurements in the rocket and satellite measurements noted above, the system descriptions (especially Dolginov et al., 1962) illustrate many of the problems in past systems. Recently Heppner and Boroson (1962) have demonstrated the feasibility of a new technique using fluxgates that is inherently simple in satellite application and is at least as accurate at near-earth field intensities as any fluxgate system known to the reviewer. Over the field range of the POGO orbit, magnetometer accuracies ranging from 20 to 60 gammas could nominally be expected. To reduce component measurement errors significantly below these figures it would probably be necessary to use a system involving the application of bias fields to an atomic resonance, optical pumping, magnetometer in a manner similar to the method described by Shapiro et al. (1960) for a vector proton magnetometer. The accuracy in a low altitude satellite system would depend on the time interval between successive bias field reversals and having negligible changes in satellite orientation within these time intervals. The system requirements are not simple and further elaboration here is not justified in view of the likely errors from items (b) and (c) in satellite systems such as POGO.

Item (b), relating magnetometer coordinates to the coordinates of attitude sensors, is a matter of mechanical rigidity which might not appear to be a fundamental problem. Consider however, that a factor of ten improvement in describing the earth's field requires an angle accuracy of about 0.1 degrees (roughly equivalent to component errors of 0 to 50 and 0 to 100 gammas depending on the component and latitude as noted in Section 2c). For this accuracy the mechanical rigidity requirement would not be severe in a highly specialized non-magnetic satellite in which the magnetometer to satellite body separation was small. However, with separations requiring long booms, such as the 6 meter booms on POGO, the rigidity problem very definitely becomes fundamental and making this source of error a fraction of the total 0.1 degree error requires either a

mechanical assembly of large weight or an auxiliary orientation sensing system relating the two coordinates systems.

Item (c), measuring the satellites orientation relative to earth coordinates, presents a major problem. There is not (to the reviewer's knowledge) any precedent in terms of satellites flown for measuring satellite orientation in all three coordinates with errors <0.1 degrees, especially under conditions that can be related to a polar orbiting satellite passing alternately in and out of the earth's shadow. For three axis data, reference to two bodies is required. Horizon and solar sensing combinations such as those in the OGO satellites (Ludwig, 1963; Scull and Ludwig, 1962) can provide measurements with errors less than 2 degrees when sensor error signals are analyzed but are inherently limited to minimum errors of 0.5 degrees by horizon variability. Looking beyond the existing art to future systems there are several possibilities, such as: (1) a satellite similar to the Orbiting Astronomical Observatory (OAO) with active, star tracking, pointing control but modified to include accurate optical measurements in the roll plane, and (2) spin stable satellites designed for negligible torque influences over several orbits and utilizing optical sensors having sharp resolution in recognizing the edge of the solar disc and a particular wavelength at the earth's horizon.

The additive nature of errors makes each of the problems noted above more severe than indicated. Solutions are in principal feasible in terms of known technological capabilities but in practice the extensive effort required to achieve vector accuracy has to be weighed against what is more easily achieved with scalar measurements. Thus when it is not obvious that vector measurements of 100 gamma accuracy possess advantages over scalar measurements accurate to several gammas, the extensive effort does not appear justified. There is not, however, an ideal substitute for direct vector measurement and ultimately such measurements will be needed for confirmation and further refinements. Vector measurements in low orbits are also needed for studies of various field variations. Thus efforts to obtain vector measurements of high accuracy can be expected in the future.

d. Magnetometers

The characteristics of optical pumping magnetometers make them particularly well suited for the WMS satellite requirements for accuracy and in-flight data storage between command readouts. To illustrate this preference for optical pumping magnetometers it is instructive to note the principal limitations of other types. Saturable core, second harmonic, magnetometers cannot be expected to meet the accuracy requirements for scalar measurements and their use for vector measurements would be very limited in satellites not having an appropriate direction reference system, as discussed in the previous section. Similarly, Hall generator and spinning coil magnetometers do not have sufficient accuracy. Proton precission, nuclear resonance, magnetometers of the free precession type, $f_p(\text{cps}) = 4257.6$ (F) with F in gauss, used in Vanguard 3 have the necessary absolute accuracy but difficult operational problems occur in data storage as a consequence of the low signal to noise ratio when a wide range of field intensities is to be measured. Solutions involve having either: (a) a high-Q selective filter that automatically adjusts to the precession frequency such that inflight frequency measurements can be made relative to a higher frequency oscillator (see Heppner et al., 1958, for an example) or (b) in-flight direct storage of the precession signal on magnetic tape for command playback. Technique (b) involves wasteful use of a high speed tape recorder which must also record a standard frequency signal such as 10 or 100 kc/s to achieve accuracy independent of variations in tape speed. Other disadvantages include: low information content per unit of power, and susceptibility to audio noise frequencies from spacecraft sources and natural whistlers and VLF emissions.

The magnetometers noted above have known operational characteristics that permit evaluation. There are other types of magnetometers to consider that fall in the category of being potentially suitable for the WMS but whose state of development does not permit use in the earliest WMS satellites. One example is the nuclear resonance, "maser" magnetometer described by Abragam et al. (1962) which utilizes the Overhauser effect between the magnetic moment of an electron

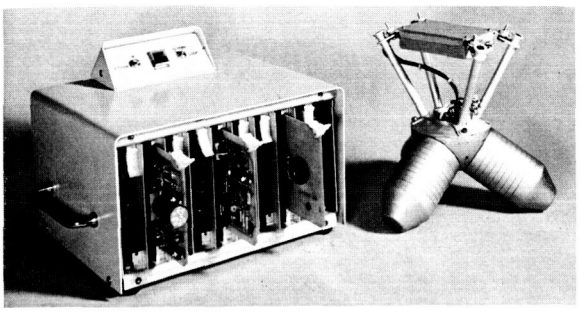


FIGURE 7.—Prototype of a three axis, omni-directional metastable helium magnetometer.

carried by a free radical and the magnetic moment of a proton to produce a continuous signal at the Larmour frequency of the proton, $f_p = 4257.6(F)$. The advantages relative to proton free precession, of greater signal strength and lower power input remove some of the operational difficulties in using proton magnetometers. A problem which has been critical, at least in the past, is obtaining samples in which the free radical lifetime permits operation for long periods of time. Another example, which shows promise as a result of observations of the electron gyrofrequency by the Topside Sounder satellite, is a magnetometer based on phase locking a transmitter to either signals received at a harmonic of the electron gyrofrequency or to a transmitting antenna impedance characteristic occurring at a harmonic of the gyrofrequency. The locked transmitter frequency, $f(\text{cps}) = N(2.8 \times 10^6)F$, with F in gauss and N designating the harmonic, can potentially provide an extremely sensitive field measurement in regions of strong magnetic fields and high electron densities like those occurring along the POGO orbit.

Returning to the subject of optical pumping, atomic resonance, magnetometers there are two types to consider: metastable helium and alkali vapor. Of the various alkalis (Li, Na, K, Rb, Cs), rubidium and cesium are the most favorable from the standpoint of the operating temperature and the resonant line width for all m levels. Within the United States developments have concentrated on rubidium. The Rb-vapor magnetometer to be flown on the first POGO satellite for the WMS is discussed in the next section. As metastable helium magnetometers may be used in future magnetic survey satellites a brief discussion is appropriate.

Figure 7 is a photograph of a prototype metastable helium magnetometer developed for the Goddard Space Flight Center by Texas Instruments, Inc. Three absorption cells, each with a circular polarizer, infrared detector, sweep coil, and cell exciting winding, are located along three orthogonal axes extending radially from a helium lamp which is driven by a 50 mc/s r–f oscillator following ignition by a high voltage. A preamplifier, to drive signals over a long boom cable, is located at the base of the lamp-sensor assembly. The lamp-sensor assembly is designed to fit in the same spherical enclosure on the POGO boom as the Rb-vapor lamp-sensor assembly discussed in the next section. Electronic units consisting of the r–f power oscillator, ignition circuits, amplifiers, servo sweep oscillator and phase lock circuits, and voltage controlled oscillator are located in the box shown. As demonstrated initially by Franken and Colegrove (1958) the Zeeman splitting (Larmor frequency separation) of a helium metastable energy level is such that the absorption of helium light near 10 830 Å occurs at different rates in the various Zeeman, m, levels. Application of a weak magnetic field at the Larmor frequency, approximately 28 cps/gamma (2.8 Mc/gauss), produces a redistribution of electron populations in the Zeeman levels and light is absorbed at this resonant frequency. Helium atoms in the metastable state are continually provided by maintaining a weak discharge in the absorption cell. In an operational magnetometer an electronic servo technique, employing a sweep oscillator and phase detection, is used such that the frequency of a voltage controlled oscillator is continuously locked to the Larmor frequency. The oscillator output frequency thus provides the measurement of scalar magnetic field intensity. Orientation effects on the field value are small, about 2 gammas with 180 degree field reversal, but require compensation for high accuracy. Signal amplitude varies approximately as $\cos^2\theta$ where θ is the angle between the optical axis and the magnetic field. This necessitates the use of three optical axes, figure 7, and signal mixing in an instrument that is to provide measurements in all orientations.

Assessment of the relative advantages and disadvantages of metastable helium and rubidium vapor magnetometers is not intended here. It is

obvious that the servo systems currently required in helium magnetometers, because of the frequency limitations of available infrared detectors, lead to a more complex electronic systen than self-oscillating Rb-vapor magnetometers. On the other hand the wider temperature range for operation of a helium magnetometer without thermal control is a distinct advantage of the helium magnetometer. Similar considerations applied to weight, power consumption, susceptibility to errors, signal to noise ratio, reliability, etc. appear with advantages and disadvantages in each type. The most important single factor in selecting a Rb-vapor magnetometer for the first POGO satellite was the state of knowledge regarding operating characteristics in a space environment within spacecraft restraints.

e. Pogo Magnetic Field Instrumentation

Descriptions of the series of satellites designated Orbiting Geophysical Observatories (OGO) which include the Eccentric Orbiting Geophysical Observatory (EGO) and the Polar Orbiting Geophysical Observatory (POGO) have been published by Scull and Ludwig (1962) and Ludwig (1963). These descriptions should be consulted for details involving the spacecraft system.

Both the EGO and POGO satellites carry Rb-vapor magnetometers with the lamp-sensor assembly enclosed in a 33 cm sphere at the end of a boom extending 6.6 meters from the main

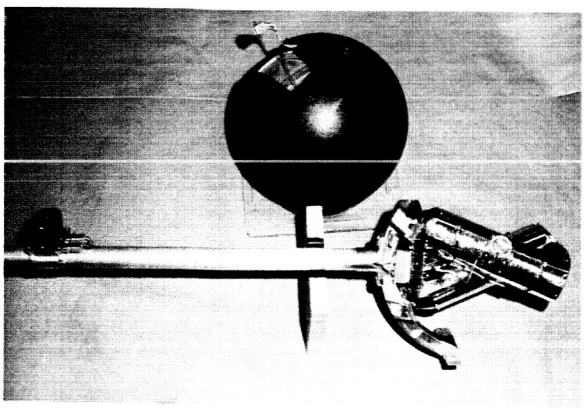


FIGURE 8.—Prototype of an EGO boom mounted rubidium vapor magnetometer consisting of two, double gas cell units. Radio frequency oscillators for lamp excitation are located on the boom to the left. The 33 cm enclosing sphere is shown removed.

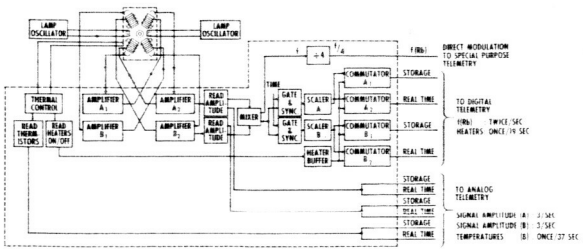


FIGURE 9.—Functional block diagram of instrumentation for POGO magnetic field experiment.

body of the spacecraft. The external appearance of the EGO and POGO units is similar with the exception that the EGO sphere includes bias field windings and also carries a 3-axis fluxgate magnetometer located about two meters closer to the spacecraft. Figure 8, a photograph of the prototype of the EGO unit, is thus also representative of the POGO unit. The block diagram in figure 9 illustrates the principal electronic units of the POGO magnetic field instrumentation.

The Rb-vapor magnetometer consists of two, self oscillating, double gas cell magnetometers which are individually similar to the type flown in Explorer 10 (Heppner et al., 1963; Ruddock, 1961). The use of two units, placed in a crossed configuration as illustrated in Figures 8 and 9, greatly reduces the size of zones of field orientation where the signal to noise ratio is low. For each unit, signal amplitude is approximately proportional to $\sin 2\theta$ where θ is the angle between the field vector and the optical axis. This gives two regions of low amplitude: a conical region centered on the optical axis ($\theta = 0°$) and a zone centered on a plane perpendicular to the optical axis ($\theta = 90°$). Thus using two such units in an "X" configuration eliminates all orientations of low amplitude except where the two $\theta = 90°$ zones intersect to form two rhombic regions. The angle between the two optical axis is 55 degrees in the case of POGO. For nominal performance the size of these regions, determined by signal to noise for reliable frequency counting, is about 500 square degrees, each, or in total about 2.5 percent of the total spherical solid angle. The 2.5 loss region could of course be eliminated using a third optical axis but the complexity is not merited using double cell units. The use of three single cells as described for helium in the previous

section, is ruled out at the field intensities along the orbits of POGO for reasons of orientation errors, as noted below.

The Larmor frequencies for the rubidium-85 and -87 isotopes in a single gas cell, self oscillating, magnetometer are given by:

$$f(\text{cps}) = 466744(F) \pm (K)359(F^2) \quad \text{for Rb-85}$$

and

$$f(\text{cps}) = 699585(F) \pm (K)216(F^2) \quad \text{for Rb-87}$$

where F is in gauss, and K, usually <0.4, indicates the fraction of the constant with the squared term that appears in a given instrument. As the F^2 dependence comes from the relative electron populations in the various m levels, which in turn depends on the field orientation relative to the optical axis, it is symmetrically + and − with respect to 180 degree field reversal. Thus by use of the double gas cell arrangement the F^2 dependence is eliminated and the field measurement is independent of orientation (see Ruddock, 1961, and Bloom, 1962 for further details). The choice of isotope, 85 or 87, is not critical. Rubidium-85 will probably be used on the first POGO for reasons of signal to noise as limited by the high frequency response of the silicon cell light detectors in fields approaching 0.65 gauss.

To decrease the probability of a single failure causing a total loss of data the design is oriented toward achieving redundancy within the restrictions of weight and power. Thus (see figure 9) a separate lamp is used in each unit rather than using one lamp for each optical axis; scaling circuits and the commutator circuits through which the count is shifted to the spacecraft data storage and telemetry system are paralleled such that the frequency is counted by alternate scalars each one-half second (this also serves as a check against any malfunction in the counting circuit that would give an incorrect reading); separate power converters, whose supplies can be individually turned on and off by ground command, are used for magnetometer units A and B and scaling lines A and B; thermal controls for A and B units are similarly separate and supplied power from the corresponding magnetometer power source. As shown in figure 9 the outputs of units A and B are mixed before entering the scaling circuits; this choice rather than treating the two outputs separately is made primarily to reduce the amount of data editing that would result from one of the two units frequently being in an orientation of low signal amplitude. The redundant features noted above have two primary consequences: (1) a failure in one of the magnetometer units will still permit data coverage of the type available from a unit with one optical axis when the faulty unit is turned off, and (2) a failure in one of the scalar-commutator units will reduce the number of measurements from 2 per second to one per second. There are, however, some types of single failures that can cause complete loss of data, both within the experiment and in the interface connections between the spacecraft and the experiment. Thus, the design is not completely redundant.

As indicated in figure 9, temperatures in the lamp-sensor assembly, the duty cycle of heater controls, and signal amplitudes are also telemetered. These functions provide information that permits detection of a malfunction in either unit which may be correctable in some cases by turning power on and off by ground command.

Digital and analog outputs are stored for playback to ground stations and also read-out in real time by the spacecraft system when in line of sight of ground stations (Scull and Ludwig, 1962; Ludwig, 1963). In addition, the Larmor frequency, $f=466744(F)$, is divided by four, to fall within a 100 kc/s modulation limit, and fed directly to the Special Purpose Telemetry transmitter where it is transmitted continuously along with four subcarrier signals from other experiments. The primary purpose of this signal is to permit studies of field fluctuations at frequencies >1 cps over receiving station sites. It also serves as a check on the correctness of the digital system through all phases: experiment, spacecraft, and surface equipment.

The sphere enclosing the lamp-sensor assembly has a black surface over all areas exposed to the sun such that its radiation properties are predictable and not subject to change through surface contamination. Aluminum coatings are used on the surfaces of the lamp-sensor assembly to minimize radiation losses from the lamp and heater circuits. Active thermal control is achieved

by having the passive, uncontrolled, temperature on the low side of the nominal 30° to 55°C operating range for the gas cells and then adding heat. Heat is added using bifilar, non-magnetic, circuits controlled by thermistor readings to several °C at the center of the temperature range Heater windings are inside the cylinders shown in figure 8. The winding that appears on the outer surface of the cylinders in figure 8 is in series opposition to the Larmor frequency feedback coil and serves to keep the feedback field from one magnetometer from appearing in the other magnetometer.

Power required for the instrumentation, including conversion losses from the 28 ± 5 volt spacecraft supply, is 6 watts with an additional 2 watts required for thermal control. The weight of the lamp-sensor assembly is 2.5 pounds. The instrumentation within the spacecraft body, including power converters and containers, weighs less than 10 pounds.

WMS DATA REDUCTION AND ANALYSIS

a. Handling of Raw Data

The POGO magnetic field survey is to yield two field measurements per second throughout the active life of the experiment and spacecraft instrumentation. This amounts to approximately 10^6 measurements per week, or 3×10^7 measurements in six months. The processing and transfer of this information from receiving station tape recordings to a computer tape suitable for analysis (see figure 10) is a major undertaking. However, in terms of the complete POGO system this magnetic field information constitutes less than 4 percent of

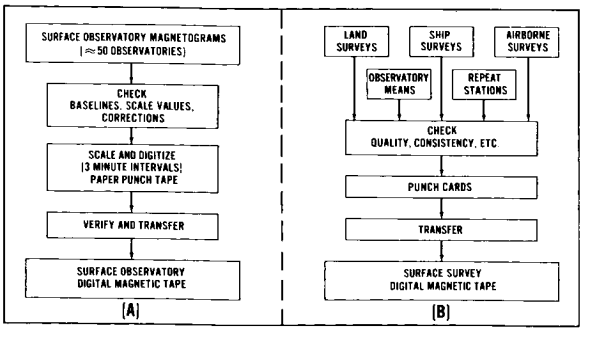


FIGURE 10.—Reduction of POGO magnetic field data from telemetry tape.

the telemetered experimental data and with the auxiliary information (see figure 9 and previous section) on signal amplitude, heater cycles, and temperatures included, the total for the experiment is less than 6 percent. With the exception of the signal transmitted via the Special Purpose Telemetry the information is digital and the processing is largely a tape handling and computer operation problem. Signals received from Special Purpose Telemetry are processed in a system particularly designed for reducing data from optical pumping magnetometers. Unlike the spacecraft digital system whose measurements are set at twice per second with a precision of ±2 cps in $466744(F)$ cps by the technique of direct frequency counting, sample periods in this system can be set in 0.001 second increments from 0.001 seconds to 10 seconds and the technique of gating a 5 Mcs. reference frequency gives a theoretical precision of ±1 part in 5×10^6 (Δt) where Δt is the set sample interval (e.g., ±0.1 gammas in a 0.5 gauss field with $\Delta t=0.1$ seconds). This provides a more than adequate check on the spacecraft digital system, but its principal use for POGO data is in studying rapid field fluctuations.

Inasmuch as analysis of the satellite data is dependent on knowing the temporal behavior of the field as recorded by surface magnetic observatories, the observatory records, which are generally in the form of 24 hour analog traces of D, H, and Z or X, Y, and Z, must also be processed to place the data in a form suitable for analysis with computers (see figure 11A). The task of converting the analog records to digital form is not particularly formidable using modern equipment but the additional attention required in checking baselines, scale values, reproducing records, verifying scalings, etc. makes it desirable to use a selected set of observatories rather than all existing observatories. Factors such as the locations of observatories and the time-space scale of equatorial and high latitude magnetic phenomena do not, however, permit selection of an ideal minimum number of observatories. Preliminary planning suggests that approximately 50 existing observatories will provide a distribution that could not be greatly improved without establishing new observatories in the large regions of the globe that are currently not represented.

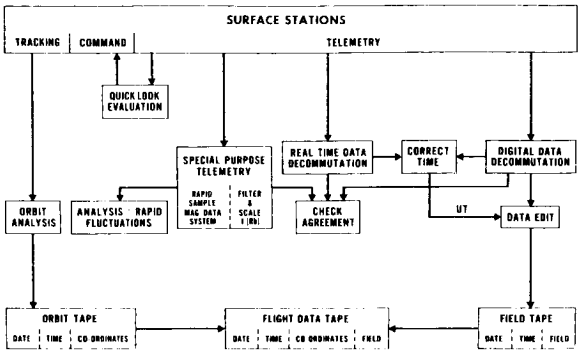


FIGURE 11.—Processing of surface magnetic field recordings and tabulated data.

Selecting a time interval between digital readings of observatory magnetograms presents additional questions for which there are not strict answers. The interval of three minutes, noted on figure 11A has been tentatively selected on the basis that it is sufficiently short to reveal most of the level changes evident on magnetograms with recording speeds of 15 or 20 mm/hr.

Utilization of surface survey data as an integral part of the analysis is also dependent on having the data in computer input form. As originally recorded most of the data from past and present surveys are in analog or tabulated form. In some cases the organization conducting the survey concerts the recordings to a digital form such as punch cards. Efficient handling of the data from diverse sources is however dependent on having all the information in a common format. Within the United States a joint effort is being made by the Coast and Geodetic Survey and the Goddard Space Flight Center to compile all available survey data as illustrated in figure 11B. These organizations with the cooperation of the U. S. National Science Foundation are also planning the the reduction of magnetograms illustrated in figure 11A.

b. Analysis Procedures

The direct application of techniques previously used in field fitting (see Section 3) to quantities of 10^6 to 10^8 measurements would not be practical even by making simplifying assumptions such as neglecting external sources and diurnal effects. When external terms are included, as they must be to get a valid description, and/or coefficients of $m=n>6$ or 10 appear probable (see Section 3c) the computation effort assumes enormous proportions in terms of computer time per coefficient. Solutions to this forthcoming problem are currently being sought along two general lines: examining how the quantity of data can be most effectively reduced with minimum loss of information, and examining alternative analysis techniques and the associated computation problems.

There are of course obvious ways of reducing the quantity of data as merely disregarding all but a small fraction of the measurements. This is hardly ideal, however, and the throwing away of data points is best done, if necessary, after first applying a selection technique having some physical basis. Various techniques of simple averaging or smoothing of data, such as applied to airborne data, are likely to increase rather than decrease errors, as a consequence of the 4 km separation of one-half second data points. One analytical approach that would serve the same function as smoothing, but is potentially more convenient in analysis, would be to represent the data along each orbit by means of a Fourier series and then reconstruct a data grid for the spherical analysis from the various Fourier series. In this and similar techniques it is essential that the time of measurement be preserved in the reduction technique such that diurnal and disturbance effects can be analyzed and removed.

The most logical approach to reducing the quantity of data for main field analysis is, in the authors opinion, one in which the data is subdivided such that: (1) measurements significantly effected by magnetic disturbances do not appear in the main field analysis, and (2) measurements are grouped according to local (solar) time such that diurnal effects can be readily recognized. When measurements are taken over periods exceeding several months a third grouping based on the dates of measurement becomes desirable for recognizing secular changes. Sub-dividing the data in this manner has additional advantages in facilitating studies directed toward understanding the source and cause of the field variations.

The flow diagram in figure 12 illustrates one method of approaching the data that fits the above criteria and makes use of the motion of the

orbit plane with time to reveal diurnal effects. The diagram is self-explanatory but several steps require comment. The selection of "quiet periods" is based on surface observatory data as indicated in figure 13. The selection, involving visual, range, and level criteria, is roughly analogous to using K_p indices along with a D_{st} (or D_0 as used here) analysis to avoid including storm recovery periods. A minimum quiet period (λ_i in figure 12) of 24 consecutive hours is essential to obtain 360 degree longitude coverage in the two local time zones sampled each orbit. Long quiet periods (e.g., more than two days) improve the density with longitude near the equator, where density is needed, but also present a greater computation problem. At this point it will be desirable to make a second selection of the data (not shown) in which measurements taken at high latitudes during polar and auroral zone disturbances are thrown out. This is the logical place to reduce the quantity of data from both the standpoint of density per unit area and disturbance effects. If the number of measurements after these steps is still too great for computation efficiency, successive selections based on eliminating successively smaller disturbance effects can be applied. A very rough estimate of the number of suitable quiet periods that might be expected is two to six such periods each month having durations ranging from one to five days. Thus, when analyzed separately (figure 12) it can be anticipated that a spherical harmonic description would be obtained for approximately each hour of local (solar) time after six months as a consequence of the orbit plane (90° inclination) moving one degree per day in local time.

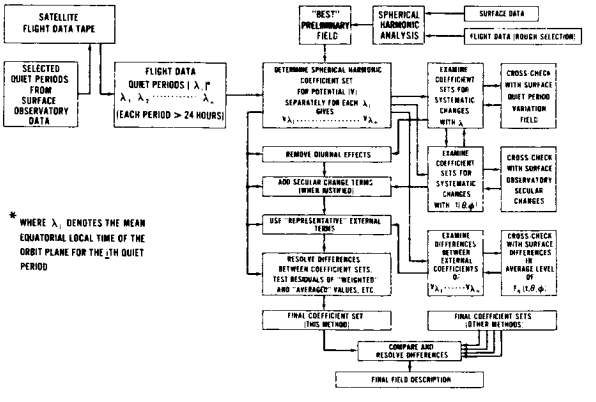


FIGURE 12.—Functional diagram illustrating steps involved in one method of arriving at field coefficients not contaminated by diurnal, disturbance, and secular effects.

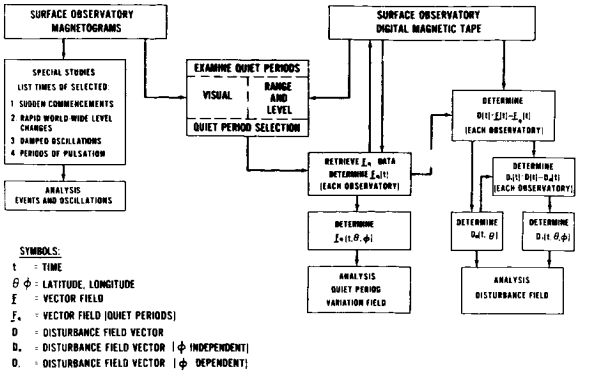


FIGURE 13.—Routine steps in analysis of surface observatory data.

The motion of the orbit plane with time, which makes λ_i equivalent to time, presents a dilemma in separating diurnal and secular effects. As the magnitude of secular change will over six months be equivalent to diurnal magnitudes in many regions of the earth, magnitude alone will not permit separation. Space will not permit full discussion here but it should be apparent that this separation will be dependent on analysis of secular change patterns at the earth's surface unless assumptions regarding the source of the diurnal variations are inserted in the analysis. Ideally one would prefer not to include such assumptions.

The difficulty and accuracy of resolving the differences between coefficient sets after removing the systematic changes depends on the magnitude of the non-systematic changes. In the case of external terms the observatory data on small level changes provides additional criteria. For the internal terms the differences may be extremely small in which case weighting and averaging according to some criteria such as mean residuals may be adequate. If the differences between sets are large the method has obviously failed.

Failure to achieve an accurate description is not likely if all the desired instrumental, spacecraft, and orbit criteria are met unless an invalid assumption is inherent in the method. The assumption of this type that is likely to cause the most

trouble is the assumption that the region of measurement is free of field sources. Over the POGO altitude range, 250 to 925 km, this would appear to be a good assumption in terms to present knowledge but this knowledge is very incomplete, especially with regard to diamagnetic contributions at magnetic latitudes >45 degrees. If it is found that the assumption of no significant sources in the region of measurement during quiet periods is not valid, modifications in the analysis procedures will be necessary.

It has also been assumed that spherical harmonic analysis is the best method for achieving the field description. There are very few examples in past works involving the direct use of the equivalent multiple dipoles or surface integrals. McNish (1940) and Lowes and Runcorn (1951) have for example used graphical methods to locate a number of dipoles for fitting the main and secular change fields. Graphical methods are not applicable where large quantities of data are involved but the direct use of multiple dipoles is not necessarily out of the question. If efficient computation techniques for analytically adjusting the dipole locations and moments are devised there are possibly other computational advantages as noted by Jensen (1962). The total problem, including external sources, has however apparently not even been approached by these methods.

POGO MEASUREMENTS OF TEMPORAL VARIATIONS

Magnetic variations of diurnal and disturbance character have been treated as undesirable noise in their relationship to the topic of this review. Their study is however an important part of the POGO experiment which should not go unmentioned. A review of the variety of field variations to be encountered in a low altitude orbit of 90 degree inclination is too large a subject to tackle here. A list of topics would include studies of: (a) the latitude-longitude distribution of the equatorial electrojet and related questions regarding the closure of the electrojet circuit and whether or not the non-equatorial S_q system is truly located in the E-region of the ionosphere, (b) the ionospheric or non-ionospheric source for the middle latitude field disturbances apparently associated with activity in the auroral zone and usually referred to as the SD and DS fields in harmonic representations (see Sugiura and Chapman, 1960), (c) the distribution of disturbance effects in auroral and polar cap regions, (d) the magnitude of diamagnetic effects associated with particle influx in auroral regions and questions regarding the existence of currents along field lines and the closure of auroral electrojet circuits, (e) the ionospheric attenuation and induction effects on hydromagnetic waves and transients as revealed by surface and satellite comparisons, etc. The analysis of surface observatory records for comparison with the field deviations seen at the satellite is an essential part of these studies. Figure 13 outlines some of the routine steps in expediting the surface analysis.

The POGO measurements because they are scalar and not vector are not ideally suited for some of these problems. They will nevertheless give a vast amount of information on causes of magnetic disturbance that is not otherwise available—particularly with regard to localized effects and ionospheric currents which cannot be studied with higher altitude satellites (Section 8).

SECULAR CHANGE AND FUTURE SURVEYS

The fact that the earth's magnetic field slowly changes in form and intensity has been recognized for several centuries and in very recent years there has been general acceptance of the belief that these changes have a hydromagnetic origin in the earth's fluid core (Elsasser, 1950; Bullard and Gellman, 1955). The attempts of various hydromagnetic theories to explain the surface patterns of secular change have given additional impetus to obtaining more accurate descriptions of these patterns. The principal impetus has come from requirements for updating magnetic charts in regions where recent measurements do not exist. For this purpose and for illustrating the patterns of change, charts showing contours of equal annual change (called isosporic charts) are constructed for each of the field elements (X, Y, and Z or D, H, and Z) and the total field (see Vestine et al., 1947b for a description of procedures).

The sparsity of observatories and repeat stations makes the construction of isoporic charts highly subjective over large areas of the earth's

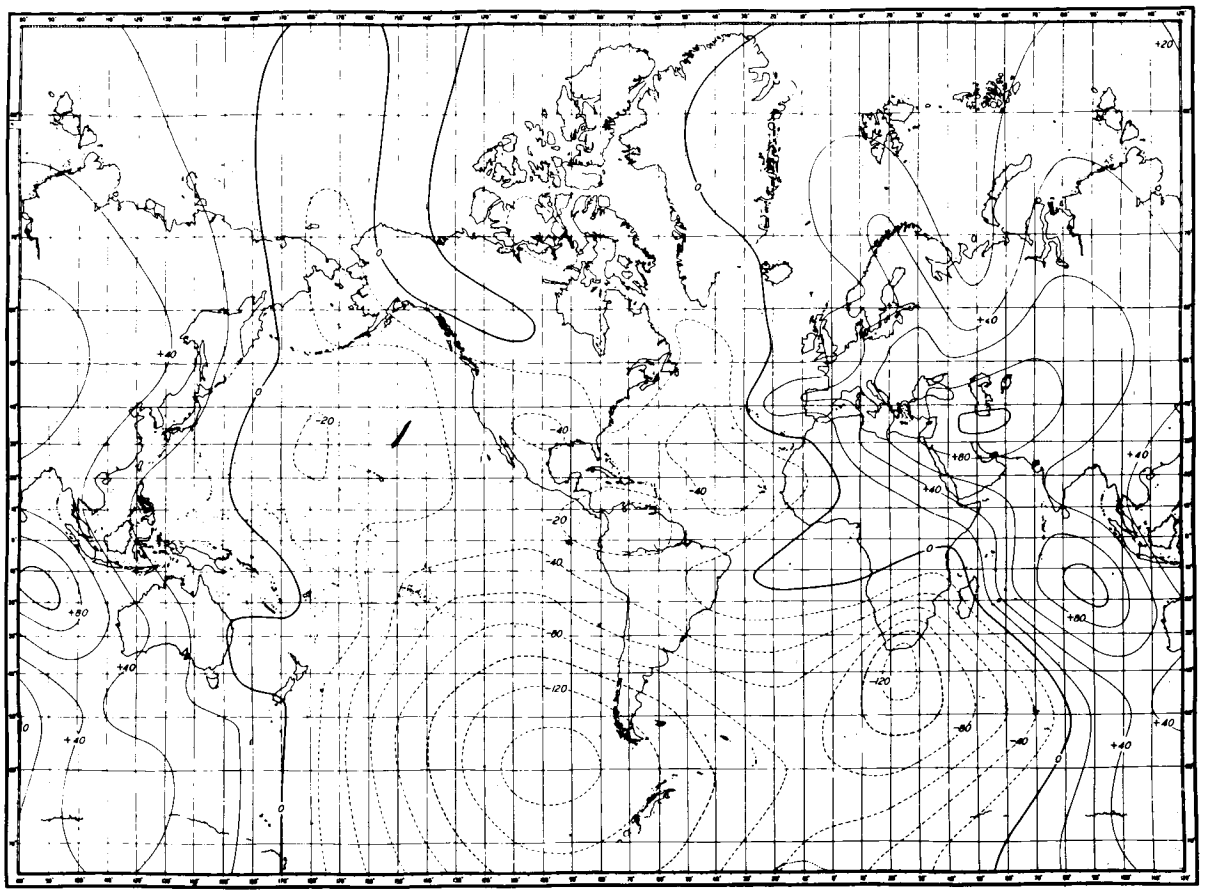


FIGURE 14.—Secular change of total magnetic field intensity in gammas per year, Epoch 1942.5 (Vestine et al., 1947b).

surface. Thus isoporic charts, such as figure 14, should not be expected to be correct in detail. Although there is reason for skepticism in using isoporic charts in many regions of the earth there has nevertheless been fair agreement in various analysis with regard to the major patterns and the movements of these patterns (especially the westward drift of the eccentric dipole) over the past 50 years (Vestine et al., 1947b; Bullard et al., 1951; Nagata and Syono, 1961; Ben Kova and Tyurmina, 1961).

Analyses are also consistent in finding that the dipole moment of the earth has in recent years been decreasing with time. The rates given by Nagata and Syono (1961) for 1955 to 1960 based on a spherical harmonic analysis of dX/dt, dY/dt, and dZ/dt are typical. These are: (a) the eccentric dipole moment is decreasing by 4.4×10^{22} emu/yr (total moment 8.05×10^{25} emu), (b) the eccentric dipole is drifting westwards by 0.3 degrees/yr, northward by 0.2 degrees/yr, and outwards by 3.4×10^{-4} earth radii/yr, and (c) other harmonics are drifting mostly westward and also at rates of several tenths of a degree per year.

Nagata and Syono (1961) and Nagata and Rikitake (1961) also point out that very intense foci appear in the isopors in the region of Antarctica with changes reaching 200 gammas/year in dZ/dt. Mansurov (1960) has also discussed the behavior in Antarctica, noted that change in the sign of dZ/dt at other locations as well is not rare, and states that superimposed periodicities of about two years were evident in Antarctica. The latter feature previously noted by Kalinin (1954) in other regions was attributed to the influence of external sources. Orlov (1961) has emphasized the need for examining solar activity effects in secular change studies.

From the magnitudes quoted above and inspection of isoporic charts such as Figure 14, it is

apparent that the POGO satellite measurements over periods of six months to a year will be influenced by average values of secular change. In regions such as those south of S. America and S. Africa (figure 14) and Antarctica the rates are such that they will not only be an influence but also should be detectable such that limited analyses can be performed. The prospects for determining the detailed world-wide patterns of secular change with follow-on satellites to POGO at intervals of three to four years are particularly good. In the regions previously mapped by Vanguard 3 in 1959 (Cain et al., 1962a; Heppner et al., 1961) there is the opportunity to study secular changes relevant to the first of the POGO satellites. These time scales for secular change studies are currently feasible only at observatories, whose distribution leaves much to be desired. The satellites should also demonstrate whether or not the present agreements in the characteristics of secular change have general validity or whether the agreements are primarily a consequence of different investigations using the same or similar data from the same observatory distribution.

SURVEYS OF THE OUTER MAGNETOSPHERE

The WMS satellite (POGO) discussed is oriented toward obtaining the best description of the earth's main field at, and close to, the earth's surface. It should reveal the existence of external field contributions and the form of the vector summation within the volume swept out by the satellite. The description does not, however, give unique solutions for the locations of the external sources just as it does not uniquely determine the locations of the internal sources. Thus one cannot necessarily expect to accurately trace the geometrical path of a field line extending to a number of earth radii even though its path up to some altitude (e.g., 1000 km if the POGO orbit is in a source free region) is well known in both hemispheres. Geomagnetic coordinates such as the L parameter for trapped particles, defined by McIlwain (1961), are similarly sensitive to external sources in weak field regions and determining the limitations will in the future rest on knowledge of the external sources.

A complete description of the earth's field in space requires direct vector measurements throughout the volume enclosed by the magnetosphere cavity formed by the solar wind. This is an immense region and any average or representative description would have to make allowance for large variations in $\Delta F/F$ with time. The existing vector measurements of Explorer 10 (Heppner et al., 1963) and Explorers 12 and 14 (respectively, Cahill and Amazeen, 1963; Cahill, 1963), the two component measurements of Explorer 6 and Pioneer 5 (Smith et al., 1960), and the scalar field measurements of the USSR Cosmic Rockets 1 and 2 (Dolginov and Puskov, 1962) have provided important data for initiating this description.

Scheduled satellites which will provide considerably more data and additional coverage within, at, and outside the magnetospheric boundary include the series of Eccentric Orbit Geophysical Observatories (EOGO) and the Interplanetary Monitor Probes (IMP). The first two satellites in each of these series are instrumented with both rubidium vapor and fluxgate magnetometers. Initial flights in each of these series precede the POGO survey. This also means that there should be more definitive knowledge regarding external sources prior to the POGO survey. This information will further facilitate the analysis of external terms in the POGO field description—a fact which was purposely neglected in the preceding survey discussions on grounds that the survey data should be independently complete for main field description.

A very significant gap in achieving a description of fields in the distant magnetosphere is evident in the fact that current satellite schedules do not include magnetic field measurements at large distances (*e.g.* 0.5 to 20 earth radii) in high latitude regions. Field mapping in these regions is of utmost importance in terms of determining the entry into the magnetosphere of solar particles and the geometry of polar cap field lines that form the anti-solar "geomagnetic tail" in response to solar wind pressures. Hopefully this situation will be rectified in the near future to complete the outer magnetosphere survey.

WMS INTERNATIONAL COOPERATION

The need for an internationally coordinated program in mapping the earth's magnetic field has been the basis for World Magnetic Survey plan-

ning (See Section 1). The recommendations put forth by international committees such as the IUGG–IAGA Committee on World Magnetic Survey and Magnetic Charts (Vestine, 1961) and the working groups for geomagnetism within COSPAR and the IQSY–CIG Committees are excellent guides for achieving coordination. They have in the past been lacking in one aspect which is particularly important to the satellite survey problem and that is the speed of transmittal of magnetograms and survey data to the World Data Centers, which were established during the IGY and are continuing operations for the WMS and IQSY periods. From the discussions in the previous sections it should be evident that surface magnetograms and survey data play an important role in the satellite data analysis. This analysis must also be conducted rapidly and efficiently and this becomes dependent on having the surface data available. In recognition of this factor the Working Group on Geomagnetism of the CIG–IQSY has recently recommended (2nd General Assembly, Rome, 1963) that observatories transmit copies of magnetograms to the World Data Centers for Geomagnetism within several weeks of the month of recording. This is a significant recommendation to groups feeling the responsibility of satellite surveys and the cooperation that will hopefully follow this recommendation will significantly contribute to the success of the WMS satellite effort.

ACKNOWLEDGMENTS

The assistance of J. C. Cain, S. Hendricks, and W. H. Farthing has been valuable in preparing this review. The author is grateful to W. Geddes of the U. S. Naval Oceanographic Office and L. R. Alldredge of the U. S. Coast and Geodetic Survey for supplying charts and analysis results pertaining to Project Magnet.

REFERENCES

Abragam, A., Combrisson, J., and Solomon, I., 'Method and Device for the Measurement of Magnetic Fields by Magnetic Resonance,' *U.S. Patent Office, Patent No.* 3,049,662, 1962.

Alldredge, L. R., and Van Voorhis, G. D., 'Depth to Sources of Magnetic Anomalies,' *J. Geophys. Res.*, **66,** 3793, 1961.

———, ———, and Davis, T. M., 'A Magnetic Profile Around the World,' *J. Geophys. Res.*, **68,** to be published, 1963.

Bauer, L. A., 'Chief Results of a Preliminary Analysis of the Earth's Magnetic Field for 1922,' *Terr. Magn.*, **28,** 1, 1923.

Benkova, N. P., and Tyurmina, L. O., 'Analytical Representation of the Geomagnetic Field over the Territory of the Soviet Union for the 1958 Epoch' (translation), *J. Geomag. and Aeron.*, **1,** 81, 1961.

Bloom, A. L., 'Principles of Operation of the Rubidium Vapor Magnetometer,' *Applied Optics*, **1,** 61, 1961.

Bullard, E. C., Freedman, C., Gellman, H., and Nixon, J., 'The Westward Drift of the Earth's Magnetic Field,' *Phil. Trans. Roy. Soc.*, A**243,** 67, 1951.

———, and Gellman, H., 'Homogeneous Dynamos and Terrestrial Magnetism,' *Trans. Roy. Soc.*, A**247,** 213, 1955.

Byrnes, B. C., 'Project Magnet,' presentation at IUGG, Helsinki, 1960; Abstr. in *Trans. Amer. Geophys. Union*, **41,** 613, 1960.

Cahill, L. J., 'Explorer Magnetometer,' Presentation at *44th Annual Meeting, Amer. Geophys. Union*, Washington, 1963.

———, and Amazeen, P. G., 'The Boundary of the Geomagnetic Field,' *J. Geophys. Res.*, **68,** 1835, 1963.

Cain, J. C. Stolarik, J. D., Shapiro, I. R., and Heppner, J. P., 'Measurements of the Geomagnetic Field by the Vanguard 3 Satellite,' NASA TN-D-1418, 1962a.

———, Shapiro, I. R., Stolarik, J. D., and Heppner, J. P., 'Vanguard 3 Magnetic Field Observations,' *J. Geophys. Res.*, **67,** 5055, 1962b.

———, and Neilon, J. R., 'Automatic Mapping of the Geomagnetic Field,' presentation at 44th Annual Meeting, *Amer. Geophys. Union*, Washington, 1963.

Chapman, S., and Bartels, J., *Geomagnetism*, Oxford Univ. Press, 1940.

———, 'Closing Address to Final Plenary Session,' CSAGI Meeting, Moscow, 1958: *Annals of the IGY* (Pergamon Press), **10,** 57, 1960.

Dolginov, S. S., Zuzgov, L. N., and Puskov, N. V., 'Preliminary Report on Geomagnetic Measurements carried out from the Third Soviet Artificial Earth Satellite,' *Artificial Earth Satellites*, **2,** Translation, Plenum Press, Inc., New York, 1960.

———, ———, and Selyutin, V. A., 'Magnetometers in the Third Soviet Earth Satellite,' *Artificial Earth Satellites*, **4,** Translation, Plenum Press, Inc., New York, 1961.

———, ———, Puskov, N. V., Tyurmina, L. O., and Fryazinov, I. V., 'Some Results of Measurements of the Constant Geomagnetic Field above the USSR from the Third Artificial Earth Satellite,' *J. Geomag. and Aeron.* (translation), **2,** 1061, 1962.

———, and Puskov, N. V., 'Some Results of the Earth's Magnetic Field Investigations in Outer Space,' COSPAR, Washington, 1962, *Space Research III*, North-Holland Publ. Co., 1962.

Elsasser, W. M., 'The Earth's Interior and Geomagnetism,' *Rev. Mod. Phys.*, **22,** 1, 1950.

FANSELAU, G. VON and KAUTZLEBEN, H., *Die Analytische Darstellung des geomagnetischen Feldes*, Geomagnetisches Institut, Potsdam, Akademie-Verlag, Berlin, 1958.

FINCH, H. F., and LEATON, B. R., 'The Earth's Main Magnetic Field-Epoch 1955,' *Mon. Not. Roy. Astron. Soc., Geophys. Suppl.*, **7**, 314, 1957.

FLEMING, J. A., 'The Earth's Magnetism and Magnetic Surveys,' *Terrestrial Magnetism and Electricity*, McGraw Hill, 1939; Dover, 1949, Chapter 1, 1939.

FOUGERE, P. F., 'Spherical Harmonic Analysis,' *J. Geophys.*, **68,** 1131, 1963.

FRANKEN, P. A., and COLEGROVE, F. D., 'Alignment of Metastable Helium Atoms by Unpolarized Resonance Radiation,' *Phys. Rev. Lett.*, **1,** 316, 1958.

HEPPNER, J. P., and BOROSON, H. R., *NASA patent disclosure* (unpublished), 1962.

———, CAIN, J. C., SHAPIRO, I. R., and STOLARIK, J. D., 'Satellite Magnetic Field Mapping,' *NASA Tech. Note* D-696; Presentation at 12th Gen. Assem. of IUGG, Helsinki, 1960(b).

———, SKILLMAN, T. L., and CAIN, J. C., 'Contributions of Rockets and Satellites to the World Magnetic Survey,' *Space Research II*, Proc. 2nd Intern. Space Sci. Sym., North-Holland Publ. Co., Amsterdam, 681, 1961.

———, STOLARIK, J. D., and MEREDITH, L. S., 'The Earth's Magnetic Field above WSPG, New Mexico from Rocket Measurements,' *J. Geophys. Res.*, **63,** 277, 1958.

———, ———, SHAPIRO, I. R., and CAIN, J. C., 'Project Vanguard Magnetic Field Instrumentation and Measurements,' *Space Research I*, Proc. First Intern. Space Sci. Sym., North-Holland Publ. Co., Amsterdam, 982, 1960a.

———, NESS, N. F., SCEARCE, C. S., and SKILLMAN, T. L., 'Explorer 10 Magnetic Field Measurements,' *J. Geophys. Res.*, **68,** 1, 1963.

IVANOV, M. M., 'Results of Magnetic Observations in Oceans during the IGY Period,' Presentation at IUGG, Helsinki, 1960; Abstract in *Trans. Amer. Geophys. Union*, **41,** 613, 1960.

———, 'The Accuracy of World Magnetic Charts for Ocean Areas,' *J. Geomag. and Aeron.* (translation), **1,** 97, 1961.

JENSEN, D. C., and WHITAKER, W. A., 'A Spherical Harmonic Analysis of the Geomagnetic Field' (Abstract), *J. Geophys. Res.*, **65,** 2500, 1960; (Report) AFSWC, Kirtland AFB, New Mexico, 1960.

———, and CAIN, J. C., 'An Interim Geomagnetic Field' (Abstract), *J. Geophys. Res.*, **67,** 3568; Corresponding Report: JENSEN, D. C., Dikewood Corporation, Albuquerque, N.M., Final Report, Contract NAS 5-1278, 1962.

KALININ, JU. D., 'The Long Period Geomagnetic Variation,' *Publ. of the Res. Inst. for Terr. Magn.*, Leningrad, 1954; Canadian DRB Translation, T379R, by HOPE, E. R., 1963.

———, 'Magnetic Observations by Artificial Sputniks and Rockets in Connection with the World Magnetic Survey,' *Space Research II*, Proc. 2nd Intern. Space Science Symposium, North-Holland Publ. Co., Amsterdam, 679, 1961.

LOWES, F. J., and RUNCORN, S. K., 'The Analysis of the Geomagnetic Secular Variation,' *Phil. Trans. Roy. Soc. A*, **243,** 525, 1951.

LUDWIG, G. H., 'The Orbiting Geophysical Observatories,' *Space Sci. Rev.*, **2,** 175, 1963.

MANSUROV, S. M., 'Geomagnetic Secular Variation in East Antarctica,' *Acad. Sci. USSR, IGY Committee, Moscow*, pp. 48–52, 1960; Canadian DRB Translation, T375R, by HOPE, E. R., 1962.

MAPLE, E., BOWEN, W. A., and SINGER, S. F., 'Measurement of the Earth's Magnetic Field at High Altitudes at White Sands, New Mexico,' *J. Geophys. Res.*, **55,** 115, 1950.

MCILWAIN, C. E., 'Coordinates for Mapping the Distribution of Magnetically Trapped Particles,' *J. Geophys. Res.*, **66,** 3681, 1961.

MCNISH, A. G., 'Physical representations of the geomagnetic field,' *Trans. Amer. Geophys. Union*, **28,** 1, 1940.

NAGATA, T., and SYONO, Y., 'Geomagnetic Secular Variation during the Period from 1955 to 1960,' *J. Geomag. and Geoelec.*, **12,** 84, 1961.

———, and RIKITAKE, T., 'Geomagnetic Secular Variation and Poloidal Magnetic Fields Produced by Convectional Motions in the Earth's Core,' *J. Geomag. and Geoelec.*, **13,** 42, 1961.

ORLOV, V. P., 'Secular Variations of the Geomagnetic Field and Methods of Using it in Compiling Small-Scale Magnetic Charts' (Translation), *J. Geomag. and Aeron.*, **1,** 226, 1961.

RUDDOCK, K. A., 'Optically Pumped Rubidium Vapor Magnetometer for Space Experiments,' *Space Research II*, Proc. 2nd Intern. Space Science Symposium, North-Holland Publ. Co., Amsterdam, 692, 1961.

SCHMIDT, A., *Tafeln der Normierten Kugelfunktionen*, Engelhard-Reyher Verlag, Gotha, 1935.

SCHONSTEDT, E. O., and IRONS, H. R., 'NOL Vector Airborne Magnetometer Type 2A,' Trans., *Amer. Geophys. Union*, **36,** 25, 1955.

SCULL, W. E., and LUDWIG, G. H., 'The Orbiting Geophysical Observatories,' *Proc. Inst. Radio Engrs.*, **50,** 2287, 1962.

SERSON, P. H., and HANNAFORD, W. L. W., 'A Statistical Analysis of Magnetic Profiles,' *J. Geophys. Res.*, **61,** 1, 1957.

———, 'Airborne Magnetic Surveys for World Charts,' Paper presented at IUGG XIIth General Assembly, Helsinki, 1960.

SHAPIRO, I. R., STOLARIK, J. D., and HEPPNER, J. P., 'The Vector Proton Magnetometer for IGY Satellite Ground Stations,' *J. Geophys. Res.*, **65,** 913, 1960.

SINGER, S. F., MAPLE, E., and BOWEN, W. A., 'Evidence for Ionospheric Currents from Rocket Experiments near the Geomagnetic Equator,' *J. Geophys. Res.*, **56,** 265, 1951.

SMITH, E. J., COLEMAN, P. J., JUDGE, D. C., and SONETT, C. P., 'Characteristics of the Extraterrestrial Current System, Explorer 6 and Pioneer 5,' *J. Geophys. Res.*, **65,** 1858, 1960.

SUGIURA, M., and CHAPMAN, S., 'The Average Morphology of Geomagnetic Storms with Sudden Commencement,' *Abh. Akad. Wiss. Gottingen, Math.-Phys. Klasse, Sonderheft,* **4,** Gottingen, 1960.

VACQUIER, V., RAFF, A. D., and WARREN, R. D., 'Horizontal Displacements in the Floor of the Northeastern Pacific Ocean,' *Geol. Soc. of Amer., Bull.*, **72,** 1251, 1961.

VESTINE, R. H., 'The Survey of the Geomagnetic Field in Space,' *Trans. Amer. Geophys. Union,* **41,** 1, 4, 1960.

VESTINE, E. H., 'Instruction Manual on World Magnetic Survey,' *IUGG, Monograph,* **11,** August 1961.

———, LANGE, I., LAPORTE, L., and SCOTT, W. E., 'The Geomagnetic Field, Its Description and Analysis,' *Carn. Inst. of Washington, Publ.*, **580,** 1947a.

———, LAPORTE, L., LANGE, I., COOPER, C., and HENDRIX, W. C., 'Description of the Earth's Main Magnetic Field and its Secular Change, 1905–1945,' *Carn. Inst. of Washington, Publ.*, **578,** 1947b.

ZMUDA, A. J., 'A Method of Analyzing Values of the Scalar Magnetic Intensity,' *J. Geophys. Res.*, **63,** 477, 1958.

EXPLORER X MAGNETIC FIELD RESULTS

J. P. HEPPNER, N. F. NESS, T. L. SKILLMAN, AND C. S. SCEARCE

Goddard Space Flight Center

Magnetic field measurements with the Explorer X satellite (1961κ) over geocentric distances of 1.8 to 42.6R_e during the period March 25–27, 1961, are presented. Analysis of the close-in data suggests the existence of a very weak ring current below 3R_e along the trajectory, but alternative explanations for the field deviations are possible. Between 8 and 22R_e, the existence of a large scale field directed away from the sun and earth within the geomagnetic cavity on the evening side of the earth is demonstrated. At distances greater than 22R_e and prior to a sudden commencement of a magnetic storm late in the flight, the geomagnetic cavity boundary crossed the satellite trajectory on six principal occasions. Interpretations bearing on the form and boundary conditions of the cavity and solar-interplanetary fields external to the cavity are presented in the light of probable explanations and alternatives. Emphasis is placed on correlations with geomagnetic activity at the earth's surface.

INTRODUCTION

Explorer X (1961κ), launched at 1517 UT, March 25, 1961, had the specific objective of obtaining vector magnetic field and plasma flux measurements along a trajectory traversing the geomagnetic field and extending into the interplanetary medium. Preliminary presentations of the magnetic field data[1,2] and the plasma data[3] were given at the International Conference on Cosmic Rays and the Earth Storm in Kyoto, Japan, September, 1961. A variety of interpretations have since been attempted with reference to various models of radiation belt currents, the geomagnetic cavity in the solar wind, the boundary properties of the cavity, and interplanetary field configurations. Direct interpretation in terms of existing models is not possible and their principal value is that of providing guidelines. In the case of model derivations of the surface geometry of the cavity, the data suggest that the model limitations stem in part from neglecting both the fields external to the cavity and the variability of the solar wind and of the external field. The uniqueness of interpretation is further limited by experimental factors such as the lack of continuous field and plasma data and of complete low energy particle spectra below 50 kev. Even more restrictive is the unavoidable fact that measurements are obtained locally along a single trajectory, whereas a satisfactory explanation of the cavity behavior necessarily involves unmeasured simultaneous behavior elsewhere. This paper summarizes major features of the magnetic field data and outlines the most plausible explanations of these features. Various alternatives are noted but are not treated in detail. Similarly, descriptions of instrumentation, calibrations, and data reduction techniques are omitted.

The magnetic field data presented at the Kyoto meeting are unchanged except for a few isolated points. The computed satellite trajectory has, however, changed considerably since the Kyoto presentation as the result of a more refined analysis of the tracking information. This demonstrated that primary use of Minitrack data led to standard deviations in position which have a relatively negligible effect on the field data. By use of the new trajectory, the large difference between computed and measured fields shown in the Kyoto paper between 1.8 and 4 earth radii (R_e) is greatly reduced. The other effect of the trajectory change is that all distances are increased at corresponding times.

SATELLITE AND PERFORMANCE

Explorer X was a 79-pound spin-stabilized satellite instrumented with: (1) a rubidium vapor magnetometer; (2) two redundant fluxgate saturable core magnetometers; (3) a plasma probe to measure the flux of low energy protons; and (4) an optical aspect system to determine the satellite's orientation relative to the earth, moon, and sun. Data were transmitted in the following sequence: Rb-magnetometer (126 sec), optical aspect (5 sec), temperature A (0.3 to 0.6 sec), fluxgate A (3 sec), temperature B (0.3 to 0.6 sec), fluxgate B (3 sec), temperature C (0.3 to 0.6 sec), and plasma probe (5 sec). The associated instrumentation included: (1) a bias field arrangement for converting the Rb-vapor magnetometer from a scalar to a vector instrument in weak fields; (2) a programmed sensitivity calibrator for the fluxgate magnetometers; (3) programming circuits for telemetry time sharing; (4) telemetry encoders for converting analog outputs to frequency; (5) a phase modulated 108-Mc transmitter; and (6) a 35-pound silver-zinc battery pack.

The active life of the satellite was estimated from battery limitations to be between 50 and 55 hours, which would provide measurements to approximately the expected apogee. In actual flight, 52 hours of calibrated data were collected and this was followed by a number of hours in which transmission continued but was unsatisfactory for data purposes. Performance was excellent except for one serious mishap: During the launching, there was excessive heating of the nose cone adjacent to the sphere containing the Rb-vapor magnetometer. The resultant outgassing caused deposition of a film on the sphere that increased the absorptivity of the surface. This raised the magnetometer temperature to 60°C after 2 hours in sunlight and caused the magnetometer to cease operating continuously. Intermittent operation continued for the next 6 hours, permitting inflight vector calibration of the fluxgates in weak fields.

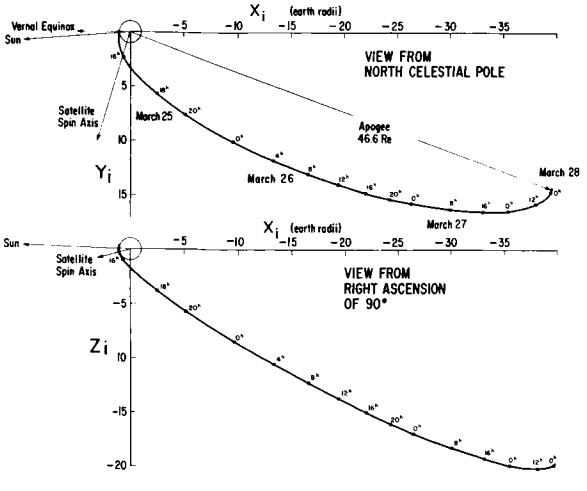


FIGURE 1.—Explorer X trajectory and satellite orientation in inertial coordinates.

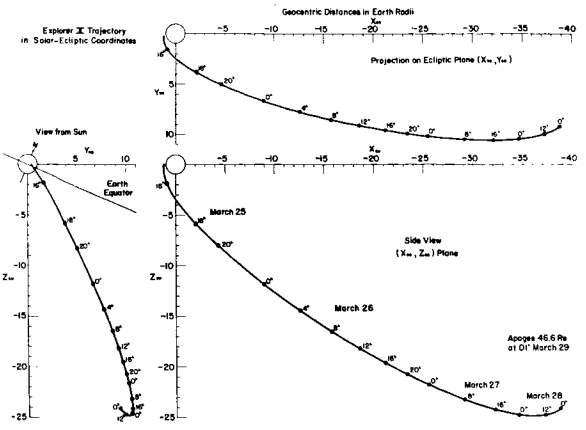


FIGURE 2.—Explorer X trajectory in solar-ecliptic coordinates.

ORBIT AND DATA COORDINATE SYSTEMS

The choice of orbit for Explorer X, relative to the earth-sun direction, was dictated by considerations of optimum geometry for plasma and optical measurements, thermal balance, radio noise background, and launch angle restrictions. The resulting orbit from launch to apogee is illustrated in figure 1 in inertial coordinates chosen so that the X_i and Z_i axes coincide, respectively, with the vernal equinox and the north celestial pole. In geographical coordinates, the orbital plane was inclined 31 degrees and oriented so that the measurements at distances 22 to $42R_e$ were located between the 2100 and 2200 local solar time meridians.

Figure 2 illustrates the orbit in solar-ecliptic coordinates chosen so that the X_{se} and Y_{se} axes lie

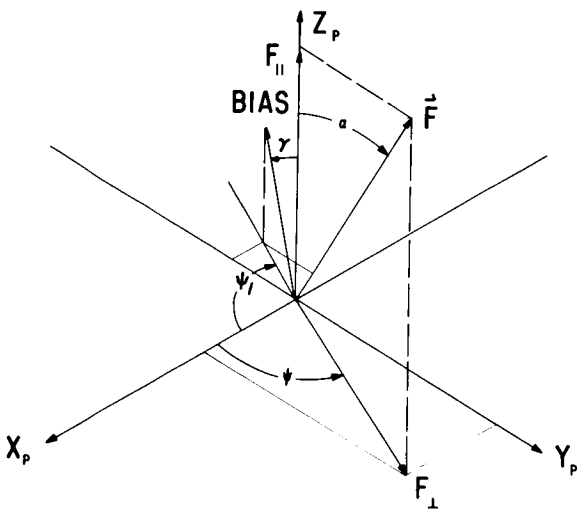


FIGURE 3.—Payload coordinate system for F, α, and ψ measurement.

in the plane of the ecliptic with the X_{se} axis pointing to the sun.

Figure 3 shows the payload coordinate system in which the data are initially reduced. The data from each fluxgate consist of the two field components (perpendicular and parallel to the spin axis) and the direction of the field in the X_pY_p plane (the angle ψ) relative to the sun. In the present paper, these coordinates are employed to $12R_e$ so that the angle ψ can be presented at

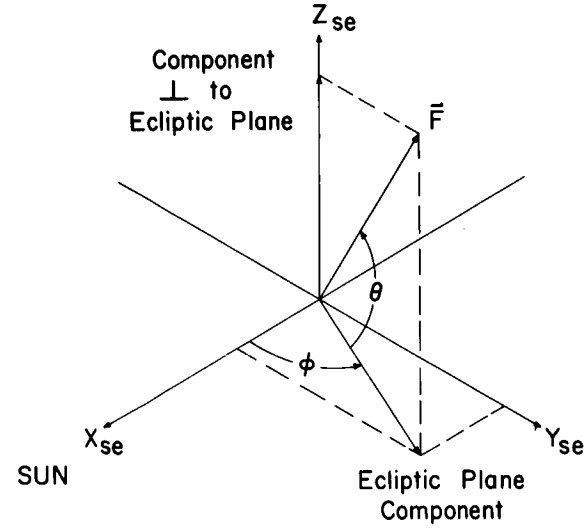


FIGURE 4.—Solar-ecliptic coordinate system for F, θ, and ϕ measurement.

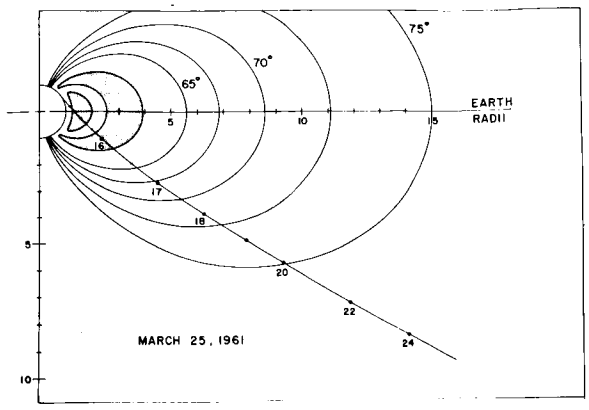


FIGURE 5.—Explorer X path in geomagnetic projection.

distances less than $7.3R_e$ where the $F_\perp$ and $F_{||}$ (the symbol F designates the total magnetic field vector) measurements are not separable and hence coordinate transformations are not possible. The spin axis Z_p is directed toward a right ascension of 71 degrees and declination of -15 degrees as is indicated in figure 1.

Figure 4 illustrates the variables used for magnetic field data following transformation to solar-ecliptic coordinates. The symbols θ and ϕ designate, respectively, solar-ecliptic latitude and longitude. Field data at distances greater than 12 R_e are presented in these coordinates.

Figure 5 illustrates the path of the trajectory relative to a projection of geomagnetic field lines

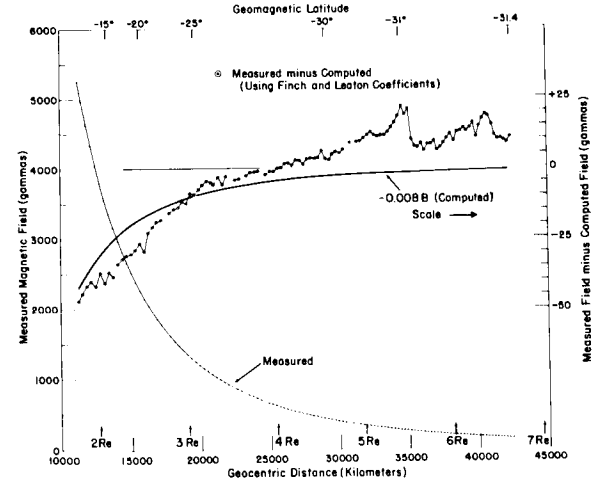


FIGURE 6.—Rb-Magnetometer scalar field measurements and field differences.

for the first nine hours of flight. Universal time is indicated at appropriate points along the trajectory.

MAGNETIC FIELD MEASUREMENTS

From 1.76 to 6.65R_e

Measurements of the total scalar field intensity for the period of continuous operation of the rubidium magnetometer are shown in figure 6. Points are plotted at one minute intervals except where the minute occurred during an encoder cycle. More detailed plotting on a scale of seconds yields a maximum fluctuation of several gammas.

To resolve spatial and/or time variation structure in the measurements, theoretical values of the field are computed by means of Finch and Leaton coefficients.[4] These are subtracted from the measured values to give the "measured minus computed" curve of figure 6. Finding the *true* difference between the measured field and the earth's main geomagnetic field depends, however, on having an independent absolute description of the main field. The true main field can only be approximated, as such a description does not exist for the measurement locations. The nearest reference data in which the accuracy of the Finch and Leaton coefficients is independently known

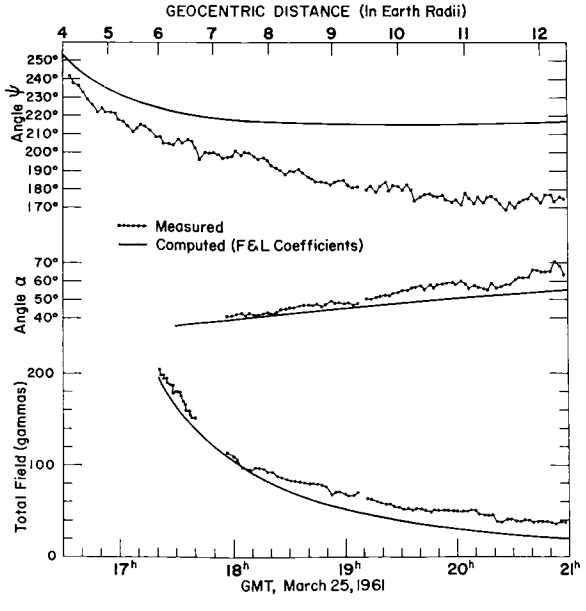


FIGURE 7.—Field measurements in payload coordinates.

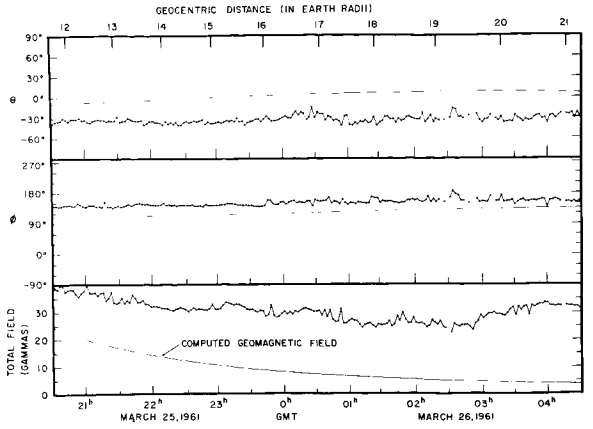


FIGURE 8.—Field measurements in solar-ecliptic coordinates.

are the Vanguard III (1959η) measurements[5] over South Africa at geocentric distances up to 10,000 km. Average errors in the Finch and Leaton values in this region are about −0.8 percent. The Explorer X measurements on Figure 6 are taken west of the Vanguard region, between Ascension Island and southwest South Africa. At the altitudes of interest, the percentage error in the computed field should not change greatly. Thus, for investigation of field differences, the computed values are multiplied by −0.8 percent and plotted on figure 6. This represents an improved reference for the observed differences. A second order improvement in this reference curve might be obtained by taking into account the probable change in error with latitude, indicated by Vanguard III, and varying the correction from −0.5 to −0.8 percent with increasing south latitude.

As was noted previously, residuals in the most recent orbit analysis would alter the computed field an insignificant amount. However, this must be treated with some reservation as there is not an absolute assurance on the accuracy of the orbital error estimate. Errors in orbit determination differentially shift the difference curves but do not alter its general shape or effect transition points (e.g., see reference 1). Thus, changes in slope of the curve, sudden changes in the difference values, and maximum and minimum percentage difference points are not appreciably effected.

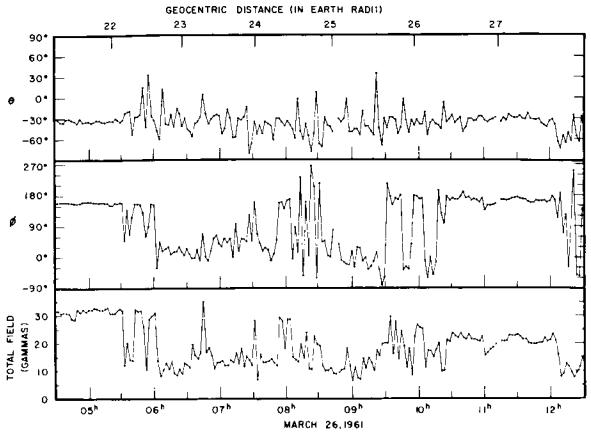


FIGURE 9.—Field measurements in solar-ecliptic coordinates.

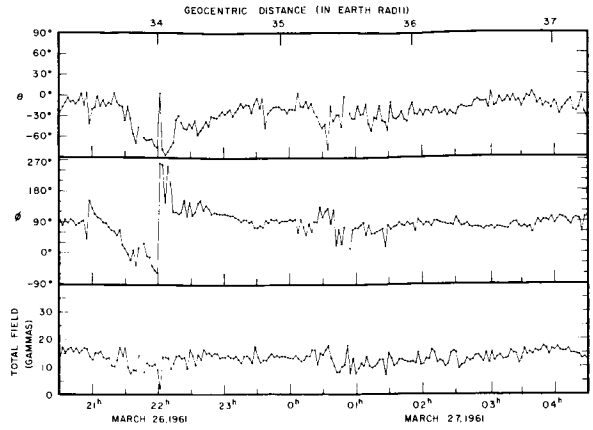


FIGURE 11.—Field measurements in solar-ecliptic coordinates.

From 7 to 12R_e

As the field intensity decreased to 100 gammas near 7.5R_e, the fluxgate magnetometers were still saturated; but intensity measurements were obtained by matching sine curves to the slopes of the readings occurring during the alternating saturation caused by the satellite spin. The accuracy of this procedure increased as the field decreased in intensity to the unsaturated value of 30 gammas, and was verified by the intermittent readings of the Rb-magnetometer. In much stronger fields, it was possible to measure ψ, the direction of the field in the X_pY_p plane of the satellite, from the time differences between the sun pulse of the aspect system and the zero crossings of the fluxgates twice per satellite spin. The accuracy of the ψ angle readings is relatively independent of field intensity. In general, the errors are estimated to be ± 3 degrees but isolated points corresponding to poor signal-to-noise conditions may be in error by 5 to 10 degrees.

The fluxgate measurements from 7.3 to 12R_e and the extension of the ψ angle values back to 4.2R_e are shown in figure 7. The total field values between 6 and 6.7R_e are included for scale continuity.

From 12 to 42R_e

Measurements from the fluxgate magnetometers at distances greater than 12R_e are shown in figures

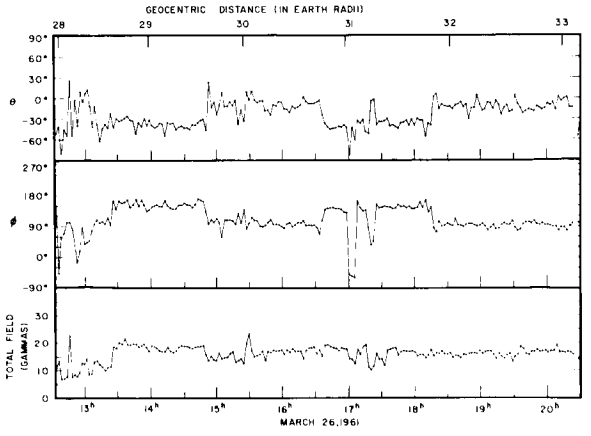


FIGURE 10.—Field measurements in solar-ecliptic coordinates.

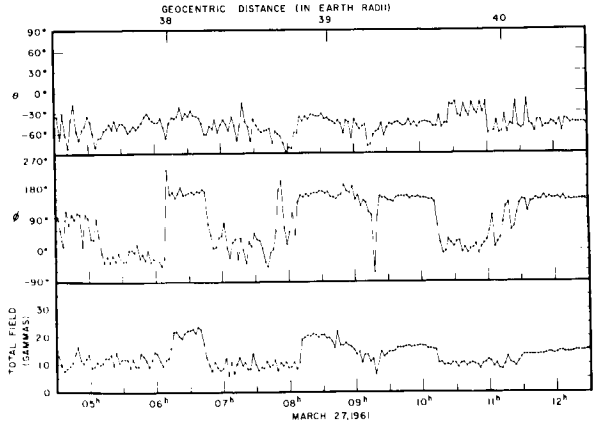


FIGURE 12.—Field measurements in solar-ecliptic coordinates.

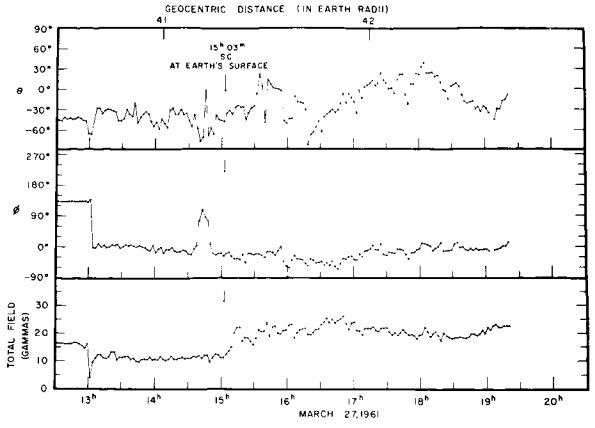


FIGURE 13.—Field measurements in solar-ecliptic coordinates.

8 through 13. Measurements between 2030 and 2100 UT near $12R_e$ are shown in both figures 7 and 8 to illustrate the change in field coordinates from F, α, and ψ (figure 3) to F, θ, and ϕ (figure 4). At this distance, $12R_e$, the field direction reaches the nearly constant value that persists until solar plasma is first detected[3,6] near $22R_e$, 0530 UT (figure 9). The computed, Finch and Leaton field is continued to $21R_e$ in θ, ϕ coordinates to illustrate the departure from the geomagnetic field.

Plotted points are averages over the fluxgate

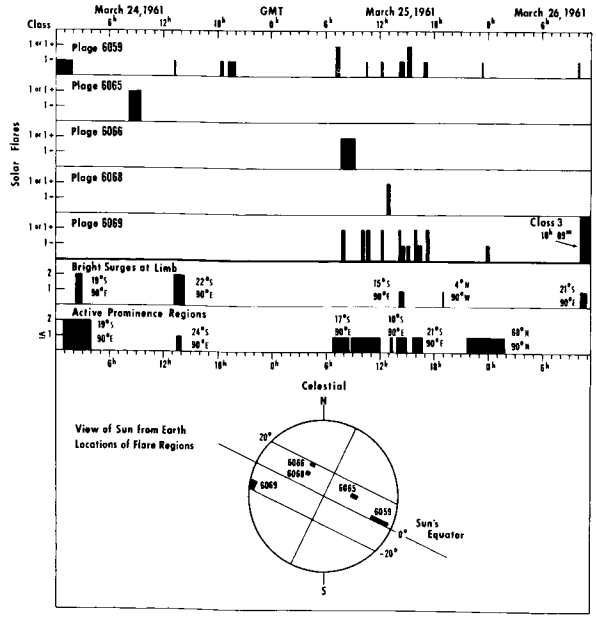


FIGURE 14.—Solar activity prior to class 3 flare.

telemetry intervals. In general, after 1330 UT, March 26, deviations within these intervals are less than several gammas. In contrast, between 0530 and 1330 UT, variations within the sample intervals having amplitudes of several gammas or more occur frequently and in selected cases variations about the average are as much as 50 percent. Rapid variations are not observed when the field magnitude is large and plasma is not detected,[3,6] e.g., between 1030 and 1200 UT, figure 9.

GENERAL SOLAR AND GEOMAGNETIC ACTIVITY

Solar observations for the period of interest were compiled by R. T. Hansen of the High

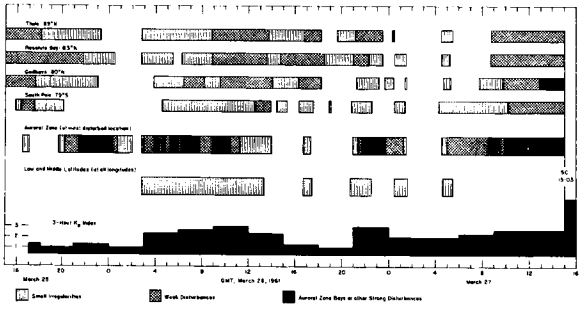


FIGURE 15.—Summary of geomagnetic activity prior to sudden commencement.

Altitude Observatory, Boulder, Colorado. Figure 14, constructed from these tabulations, summarizes the activity observed optically for two days prior to the occurrence of a Class 3 flare, at 1009 UT, March 26, whose magnetic effects were registered near the end of Explorer X's active life. Prior to the large flare, the sun cannot properly be classified as either very quiet or disturbed, and the activity level can be considered as being average. Exception to this statement might be taken on grounds that the Zürich Provisional Sunspot Number of 76, on March 24, is larger than on any day between January 5 and the flight date. However, comparison with preceding months shows that this has little significance and 76 is only 25 percent greater than the average number for the first six months of 1961. A meaningful basis for summarizing solar radio noise data has not been deduced at this writing.

An attempt to summarize geomagnetic activity prior to the sudden commencement at 1503 UT, March 27, is illustrated in figure 15. Records from 20 observatories providing fair coverage over the northern hemisphere were available for examination. In addition to the lack of south latitude stations, a gap in evaluation occurs in the auroral zone between Murmansk, USSR, and Alaska; however, observations at the Yakutsk Observatory, USSR, at a slightly lower latitude, serve as an indicator for this auroral region. The legend is relative within each of the three zones—polar cap, auroral, and low and middle latitudes. For example, the deviations classed as "small irregularities" at low latitudes are much smaller than at auroral or polar cap stations. Also, the classification of "small irregularities" at low and middle latitudes refers only to roughness or deviations on normal magnetograms that are seen by all observatories and does not include irregularities that are geographically isolated. Examples of this classification are apparent in later figures. The 3-hour K_p indices as well as the more zonal study demonstrate that conditions were neither very quiet nor disturbed on March 26 and 27 prior to the magnetic storm. The 25th, with a 24-hour K_p sum of 9, was fairly quiet. The days preceding the flight, March 21 to 24, were similar to March 26. Auroral zone magnetic activity prior to the storm was typical of average conditions. There are several intervals of 1 to 3 hours during the flight that can be classified as very quiet.

RADIATION BELT CURRENTS

Evidence for a Ring Current

By the use of preliminary trajectory analysis, it was previously concluded[1] that a field source located between the inner and outer radiation belts near the equator was necessary to explain the measurements. The more refined trajectory recently obtained has the effect of reducing the field differences from several hundred to 10 to 50 gammas, depending on the choice of reference baseline as was discussed on page 919 and illustrated in figure 6. The difference curves, however, still show a maximum percentage difference near $2.5R_e$ and a change in slope near $3R_e$. The existence of a quiet day diamagnetic ring current with maximum intensity located lower than the $3R_e$ magnetic shell along the Explorer X trajectory (or roughly $3.8R_e$ at the equator) is still possible. The magnitude of the effect is considerably less than reported for Soviet Space Probes I and II[7,8] which showed maximum effects of 800 and 140 gammas, respectively, several thousand kilometers closer to the earth than the location of the maximum intensity of the outer radiation belt. As the field shell is approximately the same as that of the Explorer X deviation, it remains to be seen if the magnitude difference can be explained on the basis that the Soviet space probes were launched during periods when the field was depressed following storm activity, as reported by Dolginov, et al.[8] Indirect evidence for a field source resembling a ring current in this region comes from the combination of Vanguard III and Explorer XII (1961ν) measurements. Under various degrees of disturbance Vanguard III[9,10] demonstrated that the source of main field depression at low latitudes must be at altitudes greater than that of maximum radiation intensity in the inner radiation belt; while, in turn, Explorer XII[11] has demonstrated that depressions in the field are not observed above the shell having an equatorial distance of approximately $4R_e$. This restricts the possible locations of a ring current maximum to roughly 1.5 to $4R_e$ under disturbed conditions and to less than $4R_e$ during quiet periods. Explorer XII[12] also revealed the existence of an intense proton flux in the outer belt, the peak intensity occurring near $3.5R_e$. The fact that the proton flux increased in the same region during a magnetic storm[13] suggests that the observed protons, which had energies $E>100$ kev, may contribute significantly to ring current effects under both quiet and disturbed conditions. The possibility that an equal or greater contribution to ring current effects may come from lower energy protons in a greater number density cannot be ignored until energies below 100 kev are measured. At much lower energies, the Explorer X plasma measurements suggest that the energy spectra may be discontinuous with distance in this region. This indication comes from the disappearance above $3.0R_e$ of a strong spin modulated signal that persisted from 1.3 to $3.0R_e$.[3] Rossi[14] and H. S. Bridge (personal communica-

tion) have interpreted this as velocity "scooping" of a stationary plasma which was terminated by positive charge accumulation on the satellite. An explanation has not been given for the double, rather than single, peak modulation which occurred on the last data sample, at $3.0R_e$, prior to the disappearance.[14] The possible relationship between the disappearance of the plasma signal and the ring current, if not attributed to satellite charging, is only indirect in suggesting sharp changes in energy spectra with distance at low energies. The relationship to the field angle changes noted in the following section may be more direct.

At distances greater than $4.8R_e$, small total field deviations (figure 6) are observed that do not have an apparent coincidence with time variations at the earth's surface. Instead, they may represent local variations in the field. The sign of the largest variations, about 15 gammas, is positive and thus does not suggest a ring current above $5R_e$ such as do the strong, quiet day currents between 6 and $10R_e$ derived from Explorer VI (1961δ) observations.[15]

Evidence for Meridional Currents

Between 1.5 and $3.2R_e$, the angle ψ was measurable by the technique described on page 920 for distances greater than $4.2R_e$. Between 3.2 and $4.2R_e$, however, the fluxgate saturation was in one direction and ψ could not be measured. Figure 16 shows the ψ angle measurements between 1.5 and $3.2R_e$ along with a computed curve of the ψ angle expected on the basis of a Finch and Leaton field.

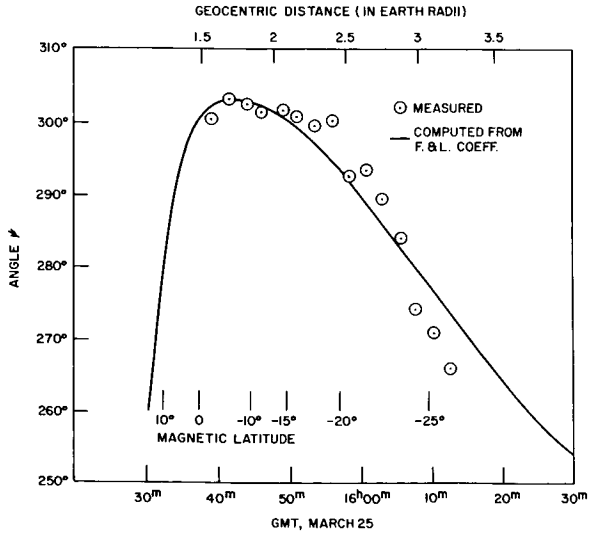


FIGURE 16.—Angle ψ changes at low altitude (see text).

The interesting feature of these measurements is that angle deviations, thought to be greater than reading errors, occur near $2.5R_e$ and $3R_e$ in coincidence with other data features independently noted. For example: (1) the ψ angle point that falls on the curve (figure 16, near $2.5R_e$) coincides in time with the small dip in the total field difference curve near $2.5R_e$ in figure 6: (2) the first point that falls below the computed curve at $2.99R_e$ coincides with the occurrence of two peaks per satellite roll in the plasma observations prior to the disappearance of plasma signal (see above); and (3) the ψ angle shift near $3R_e$ also coincides approximately with the change in slope of the total field difference curve in figure 6. If it assumed that the ψ angle changes in strong fields are significant, an explanation must be sought. In the 2.5 to $3R_e$ region, a shift in the actual ψ angle relative to the computed ψ angle is equivalent to a change in declination: that is, a decrease in ψ corresponds to an increase in east declination of the magnetic field meridian plane. This follows from the consideration that a sizeable change in total field intensity does not accompany the angle change. Changing the ψ angle by 5 to 10 degrees is equivalent to adding an east-west vector of about 120 to 240 gammas at $3R_e$ and 180 to 360 gammas at $2.5R_e$. The change must also be of local origin to explain the abruptness.

These changes, unless related to an unknown boundary condition, are not of the type expected from a ring current source and thus do not necessarily support the ring current arguments given earlier. Geometrically, the angle changes are most easily explained by postulating electric currents along field lines, a condition which violates the usual assumption that the field lines are also equipotential lines due to the extremely high electrical conductivity along the lines above the ionosphere. This condition may, however, be violated if: (1) the field lines intersect the ionosphere at points in opposite hemispheres which are not at the same potential: and (2) there is an ionospheric mechanism, equivalent to a generator in one or both hemispheres, that

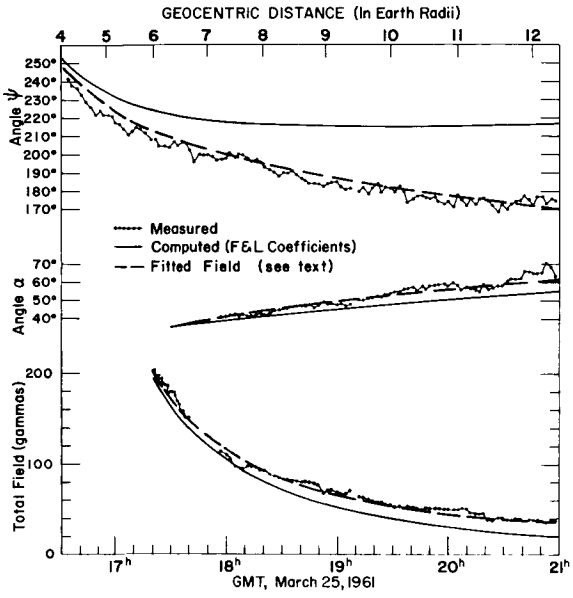


FIGURE 17.—Field fit obtained by adding a constant vector (4–12 R_e).

continuously maintains a small potential difference between the conjugate points. A difficulty in this explanation comes from the lack of observational evidence for the completion of the circuit in the ionosphere on a more localized scale than an ionospheric S_q system.

The above can only be treated as speculative; however, the ψ angle changes appear significant enough to merit future investigation.

Effects of the Geomagnetic Cavity at Several R_e

The total field deviations between 1.76 and $3R_e$ were interpreted in a preceding section in terms of a current source within the same region. An alternative explanation for the deviations is suggested by the extension inward of a superimposed field presumably caused by the compression of the geomagnetic field by the solar wind. Thus, the magnetic field in the interior of the geomagnetic cavity can be considered as a superposition of the undisturbed geomagnetic field and the field due to sheath currents flowing on the cavity boundary. Fitting the observations by vector addition of an anomalous field to the geomagnetic field should thus reveal characteristics of the cavity field along the flight trajectory.

At distances above $7R_e$, where complete vector data are available, and below $22R_e$ the superimposed field yielding an accurate fit to the observations is approximately constant with a magnitude 25 to 30 gammas and directed away from the sun and below the ecliptic plane. Indeed, for the interval between 4 and $12R_e$ shown in figure 17 a constant vector of 25.5 gammas directed so that $\phi = 182$ degrees and $\theta = -51$ degrees yields an excellent fit. At distances greater than $12R_e$ (figure 8) a fit to the data is obtained by gradually increasing the magnitude to 30 gammas and rotating the vector so that ϕ is decreased and θ is less negative.

The fit obtained with a nearly constant vector suggests continuing it back along the trajectory to investigate its effect on the close-in magnitude data. In figure 18 this is illustrated for the range 1.76 to $6.7R_e$, where the vector used is the same as was used in figure 17, and the measured field is subtracted from the resultant superposition field. The excellent fit above $3R_e$ is apparent. Between 1.76 and $3R_e$, the fit becomes poorer but, nevertheless, shows the feature of a change from a positive to a negative anomaly. This suggests that it may be unnecessary to explain the deviations below $3R_e$ in terms of a diamagnetic ring current. It suggests further that quiet day ring currents deduced from Explorer VI component data[15] should be re-examined with assumptions other than the preservation of magnetic meridian

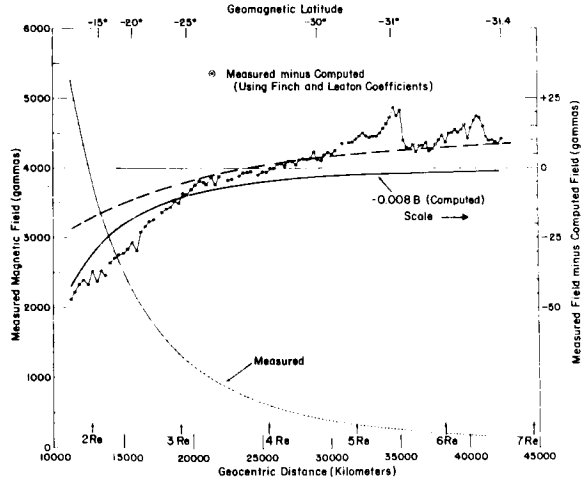


FIGURE 18.—Field fit obtained by adding a constant vector (2–6 R_e).

planes. Currently, this is difficult, as theoretical models of the cavity treat only the surface shape and do not predict the field interior to the cavity.

An uncertainty in the above arguments arises in that the same superposition vector when extrapolated to the surface of the earth would produce diurnal changes which are not entirely consistent with surface observations. This makes it apparent that the effects of a field due to cavity surface currents become more complicated near the earth. In turn, it raises a question as to how close to the earth the extrapolation discussed above is valid.

MAGNETIC FIELD AND PLASMA CORRELATIONS

For a detailed account of the simultaneous plasma measurements conducted by the MIT group, the reader should consult the papers by Bridge, Dilworth, et al.[3] and Bonetti, Bridge, et al.[6] Principal features to note are the following:

1. Between approximately 3 and $22R_e$ (or at 0530 UT, March 26), plasma was not observed with the exception of a very weak signal near 0425 UT, March 26. The direction of the magnetic field at distances 12 and $22R_e$ (figures 8 and 9) correspondingly remained nearly constant.

2. At 0533 UT, March 26, the field angles and magnitude changed abruptly, and, on subsequent samples, plasma was observed whenever the 0 to 250, 0 to 800, and 0 to 2300 volt measurements were coincident with magnetic field angles substantially different than those observed between 12 and $22R_e$.

3. Between 0533 UT, March 26 (figure 9), and the time of the sudden commencement at 1503 UT, March 27 (figure 13), there were a number of periods in which plasma was not observed and the magnetic field direction returned approximately to the direction noted between 12 and $22R_e$. In total, these periods occupy about 25 percent of the measurement time between 0533 UT, March 26 and 1503 UT, March 27.

Correlations are also noted in detail. For example, on Figure 9 near 0800 UT the field returns briefly to values similar to those observed before 0530 UT and plasma signal disappears. Also, on figure 9 near 1100 UT, the deflection of the field away from a stable condition coincides with an isolated reappearance of plasma. A similar example occurs during the field deviations between 0900 and 0930 UT, March 27.

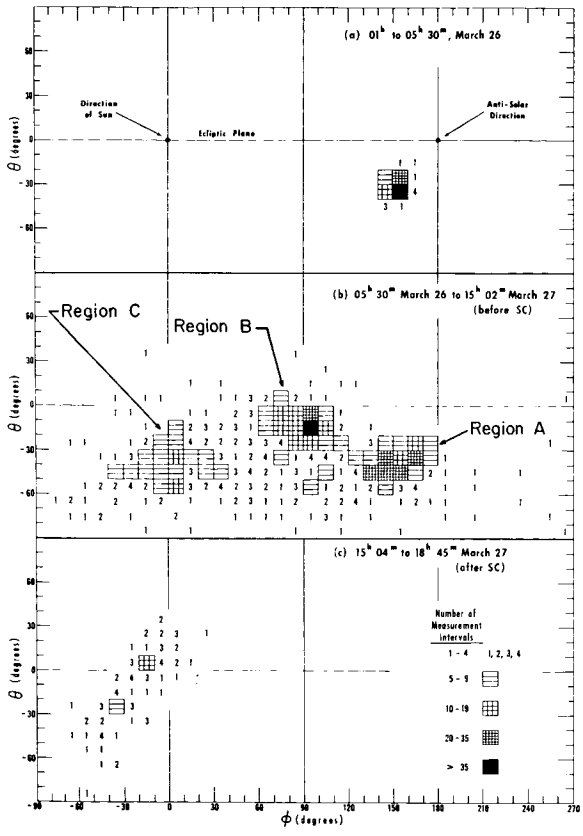


FIGURE 19.—Statistical distribution of orientation of the field vector.

In general, it can be noted that during periods where large changes in the field occur between successive readings there is also more relative variation in the plasma spectra and total flux. In the following sections other features of the correlation are noted.

CHARACTERISTIC FIELD VECTORS AND SIGNIFICANT CORRELATIONS

Average Values

From inspection of the angles θ and ϕ as a function of time (figures 9 through 13), it is apparent that a given orientation of the field vector that persists over a number of readings tends to reoccur at later times. It is also apparent that a given range of orientations frequently persists for fractions of an hour to 2.5 hours

without large changes. Figure 19 illustrates the distribution of field orientations in a θ, ϕ plot constructed by noting the number of measurements occurring in each 10×10 degree sector. As anticipated from inspection of figures 9 through 13, orientations of the field vector between 0530 UT, March 26, and 1502 UT, March 27 (figure 19b), group preferentially into three regions which, for convenience, are designated A, B, and C.

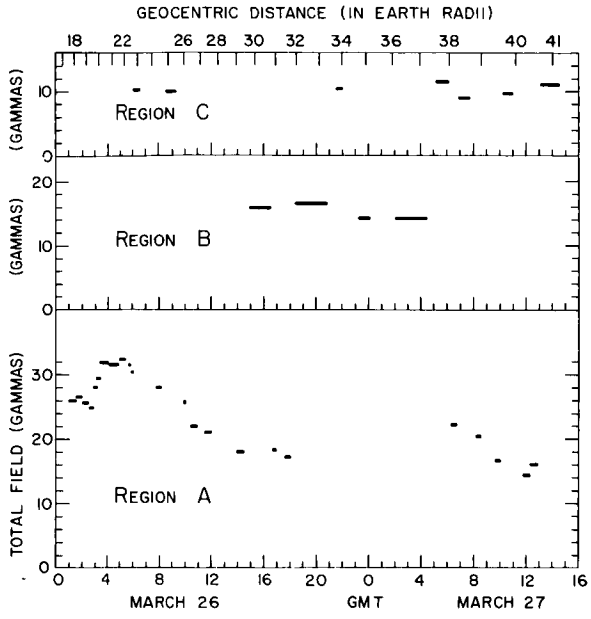


FIGURE 20.—Average field intensities grouped according to characteristic field orientations.

Considering only those measurements that occur in a sequence of measurements in which the field orientation remains in one of the three regions A, B, and C, corresponding field intensities are obtained by averaging the values within each sequence. Figure 20 shows these averages. The length of the individual bars gives the time interval over which the intensities were averaged. From figure 20, it is apparent that the level of field intensity is distinctly different in Regions B and C, and that field intensities in Region A appear to vary systematically with time.

Region A Characteristics

A large fraction of the measurements contributing to the density of Region A (figure 19b) occurs

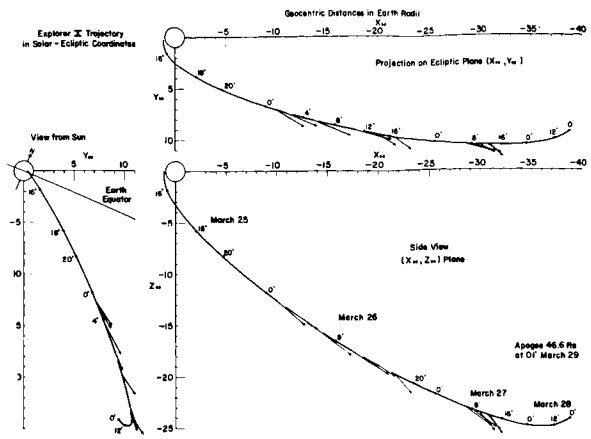


FIGURE 21.—Average field vectors during region A periods.

during periods when solar plasma is not observed. The orientations of the field at these times closely resembles the orientation prior to the first appearance of solar plasma near 0530 UT, March 26, as illustrated in figures 19a and 21. Thus, it is logical to postulate that the fields measured in the absence of plasma at distances greater than $22R_e$ are extensions of the field observed between 12 and $22R_e$. If this field is in turn attributed to compression and draping of the geomagnetic field by the solar wind, it follows that the satellite was inside the geomagnetic cavity during the Region A periods (figures 20 and 21). Acceptance of this argument depends, however, on also showing that the magnitude of the Region A field decreases with distance from the earth. As the next section shows, this condition can be satisfied on the basis

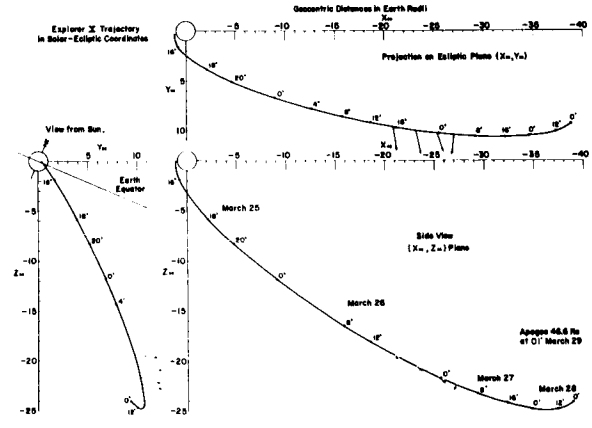


FIGURE 22.—Average field vectors during region B periods.

of assumptions that can be at least subjectively supported.

Region B Characteristics

The concentration of measurements in Region B of Figure 19b comes primarily from four periods occurring between 1500 UT, March 26, and 0430 UT, March 27 (figures 20 and 22). A number of conditions appear to correlate with the Region B field direction:

1. Plasma is always observed.
2. The plasma flux and energy spectra appears to be less variable during these periods than at other times.
3. Magnetic activity on the surface of the earth during these periods was significantly lower than the average activity after 0300 UT, March 26. This is particularly obvious, from figure 15, for the two periods of over two hours duration near 1900 UT, March 26, and 0400 UT, March 27.
4. Three of the four periods occur within the 12 hours, approximately 1800 UT, March 26, to 0600 UT, March 27, in which plasma is continually present and there is not a recurrence of an interval having Region A characteristics.

The relative importance of (3) and (4) above in interpretation is uncertain. The entire 12-hour period in which the Region A, or cavity, field condition never appears can be treated as a single interval in which the Region B field configuration is disrupted for periods of 1.5 to 2 hours, starting near 2100, 0000, and 0430 UT in coincidence with field changes at the earth's surface (See page 929). As discussed on pages 927 and 928, it also appears that during the 12-hour interval, the geomagnetic cavity may be shifted away from the satellite as a consequence of the diurnal change in the position of the dipole axis relative to the solar stream. Regardless of their relative importance the two characteristics (3) and (4) mutually support the concept that the fields measured during these periods were of solar-interplanetary origin. The degree to which the proximity of the geomagnetic cavity affects the field configuration is not determinable, but the relative stability of θ and ϕ, especially during the longest periods, suggests that the effect may be small and perhaps negligible.

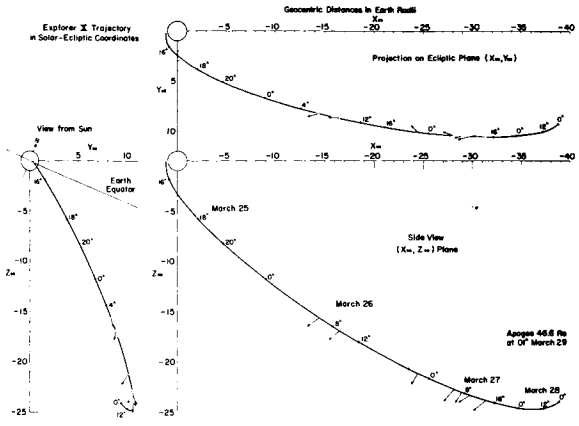


FIGURE 23.—Average field vectors during region C periods.

Region C Characteristics

When the field is oriented in the general direction of Region C it is usually more variable between successive measurements than in the case of Regions A and B. In Figure 19b, the absence of a dense concentration of measurements in any 10 degree square and the general spread of angles makes this apparent. Region C is nevertheless distinct and the fact that a statistical grouping exists comes largely from the recurrence of a number of periods in which the field has this general direction for 0.5 to 1.5 hours. The following conditions are characteristic for these periods:

1. The average field intensity is lower than during other periods (see figure 20).
2. Plasma is always observed.
3. Magnetic activity on the earth's surface is greater than during periods of Region B field orientation.
4. The periods tend to alternate in occurrence with periods of Region A field orientation.

The exception to item 4, between 2100 and 2130 UT, March 26, gives added support to the significance of item 3, which will be discussed in more detail shortly. The field direction and the fact that field intensity during these periods is independent of distance from the earth (figure 20) support the view that the field originates external to the geomagnetic cavity. It is possible, however, that the field configuration may be influenced considerably by the close proximity of the geomagnetic cavity.

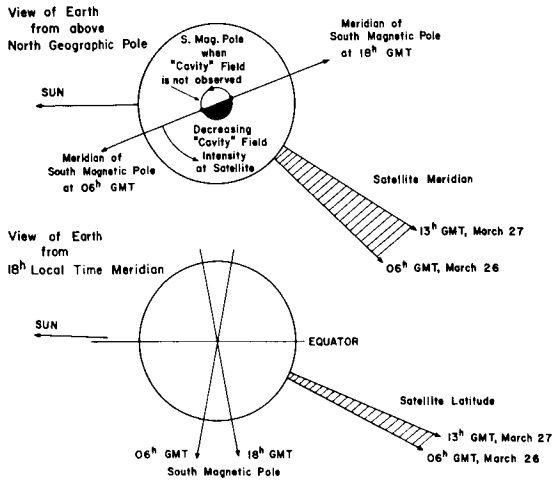


FIGURE 24.—Relative positions of the sun, dipole axis, and satellite when region A, or cavity, field is observed.

FIELD INTENSITY VERSUS DISTANCE (REGION A)

As was noted in the last section, the argument that the fields measured when plasma is not observed are of geomagnetic origin, based primarily on the direction of the field, must also be supported by evidence that the field intensity decreases with distance from the earth. Considering first the range 11.5 to $22R_e$ (figures 7, 8, and 9) it is apparent that the field decreases much less rapidly than the computed field. It is also apparent that points can be chosen (e.g., at 16 and $21R_e$) so that $F(r)=F(r_1)(r_1/r)\chi$ with $r>r_1$ gives a power $\chi=0$ and a dependence is not found. Thus, the approach of looking for the maximum dependence (i.e., maximum χ) that can be partially justified is adopted. On this basis, the higher intensities following 0254 UT, March 26, are neglected on several possible grounds: (1) that these coincide with an increase in geomagnetic activity (figure 15) having the character of greater compression of the earth's field (see page 929); (2) that the higher intensity may be a diurnal effect, explained later; and (3) that the higher intensity may be a cavity boundary effect. Using average values near 16.3 and $19.4R_e$ relative to $11.7R_e$ yields $\chi=0.85$.

At distances above $22R_e$, it is again apparent that $\chi=0$ and even a negative χ can be found from the 27 to $32R_e$ periods relative to the 38 to $39R_e$ periods (see figure 20, Region A).

If, however, a more general view of the time sequence of occurrences of Region A fields is considered, there is reason to believe that the observed intensity has diurnal characteristics. The evidence comes from the decreasing level of intensity from 0600 to 1800 UT, March 26, and from 0600 to 1200 UT, March 27, as well as the absence of Region A fields from 1800 to 0600 UT. As a diurnal effect would most likely appear as a result of changes in the relative positions of the solar stream, the earth's dipole, and the satellite, these positions are illustrated in figure 24. The geometry indicates that the field intensity at the satellite when it is inside the cavity may depend on the intensity of the field along the satellite-sun line on the side of the earth facing the sun. This suggests the calculation of χ on the basis of Region A intervals separated by approximately 24 hours. Using various combinations of intervals on March 26 and 27, separated by approximately 22 to 26 hours, gives the range $\chi=0.2$ to 1.0 with most values between $\chi=0.6$ and 1.0. Similarly, using minimum average intensities near $40.3R_e$ relative to intensities at $11.7R_e$ gives $\chi=0.81$.

The above leads to a generalization that the average decrease in field intensity with distance inside the geomagnetic cavity along the Explorer X trajectory can be represented by $F(r)=F(r_1)(r_1/r)^\chi$ with $0.6<\chi<1.0$ for $11.5R_e<r<40.5R_e$ and $r>r_1$. Different assumptions, or criticism of the assumptions used, can lead to a smaller χ (i.e., less dependence) but it seems unlikely that assumptions giving $\chi>1.0$ can be justified.

CORRELATIONS WITH SURFACE GEOMAGNETIC VARIATIONS

Prestorm Correlations

The general level of geomagnetic activity during the flight was illustrated in figure 15 and discussed on pages 921 through 922. A complete discussion of possible detailed correlations between satellite and surface measurements is beyond the scope of this paper; thus, the following discussion is limited to several examples involving primarily low latitude activity.

The period 0245 to 1345 UT, March 26, is of particular interest because of the simultaneous and similar behavior of all the low latitude magnetograms available for examination. Between 0245 and 0300 UT, March 26, the horizontal component increases from a previously quiet condition and remains positive with irregular but slow variations that are typically 5 or 10 gammas in amplitude for a number of hours. This activity disappears gradually between 1300 and 1345 UT, March 26. At much higher frequencies within this period the micropulsation energy spectra near 0.1 cps at Fredericksburg, Va., shifted to higher frequencies near 0600 UT,[2] and micropulsations of "pearl" character were observed in California (L. R. Tepley, private communication) and in the USSR (V. H. Troitskaya, informal communication at Kyoto meeting September, 1961) for several hours after 0600 UT. Through this same time interval the field intensity at the satellite increased near 0300 UT, March 26, and remained at a high level until solar plasma was observed at 0530 UT. Between 0530 and 1330 UT (figures 9 and 10), there was noticeably more variation between successive measurements by the satellite than at other times and much larger variations within the 6-second measuring period, as noted on page 920 and references 1 and 2. If the low latitude $+\Delta H$ irregularities are attributed to increased and more irregular compression of the geomagnetic field by the solar stream, it follows that the cavity dimensions and geometry should also vary. This may explain the short intervals of plasma disappearance and Region A fields between 0530 and 1030 UT, March 26. Similarly, change in the solar wind and/or any cavity influence on the configuration of external fields may explain the variability of the external field during this period. The average plasma flux[3,6] when present during this period appears to be higher than during most periods prior to the magnetic storm.

A sequence of three isolated cases of $+\Delta H$ roughness in surface magnetograms occurs during the 12-hour period in which plasma appears to be continually present and stable Region A field conditions do not appear. In figure 25 an attempt is made to illustrate the simultaneous surface and satellite measurements. Details in the satellite measurements can be seen more clearly by

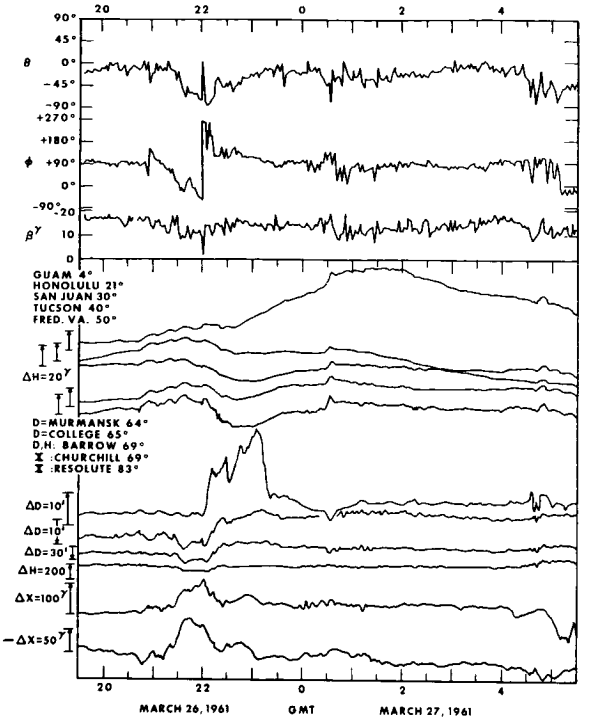


FIGURE 25.—Simultaneous surface and satellite measurements.

inspection of figures 10, 11, and 12. The low latitude times of interest are: (1) approximately 2045 and 2245 UT on March 26; (2) the change near 0030 UT on March 27; and (3) the change near 0445 UT on March 27. The first of these coincides with a 2-hour period at the satellite in which the field changes and rotates away from a relatively stable Region B configuration during the first hour, and rotates back during the second hours, as discussed in more detail below. The second and the irregularities that follow, coincide with a period of irregular variations at the satellite between two intervals in which the field has the Region B configuration. The third coincides approximately with an irregular slow transition from a Region B to Region C field configuration at the satellite. In both the second and third examples, the variations appear to begin at the satellite prior to the low latitude changes and in closer coincidence with small changes at high latitudes (e.g., the fluctuations starting near 0000 UT at Churchill and Pt. Barrow and near 0430 UT at most high latitude stations). The maximum

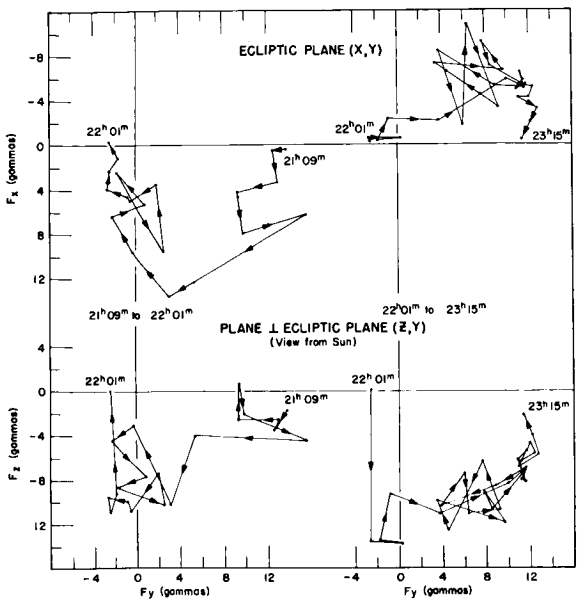


FIGURE 26.—Rotation of the field vector in solar-ecliptic coordinates between 2109 and 2315 U.T., March 26. Arrow marks indicate the time sequence of vector end points.

variation at the satellite occurs, however, more in coincidence with, or following slightly, the low latitude changes. In case (1), the low latitude initial change appears to occur several minutes before the change at the satellite, depending on point selection, and an initial sharp change at high latitudes is not readily identified at a large number of stations.

The behavior of the field vector at the satellite during the case (1) interval is particularly interesting. In figure 26, the successive positions of the end point of the vector is shown in solar-ecliptic coordinates for the center portion of this interval. After 2109 UT, the field rotated irregularly to the Region C field direction and at 2201 UT abruptly dropped in intensity to 2.5 gammas with a corresponding abrupt change in direction. Following the measurement at 2201 UT, the intensity and direction abruptly changed back to roughly the previous intensity level but to a different orientation from which it then rotated to a direction not characterized by Region A, B, or C. It then gradually returned to the Region B orientation. In addition to the total time interval correlation at low latitudes, the sharp change at 2201 UT may be significantly related to changes at high latitude. The sharp, simultaneous change at Murmansk on the night side of the earth is obvious in figure 25. On the dayside of the earth, near local noon in Alaska, the declination changes considerably with only slight changes in field intensity. The most striking feature, however, is that the level of the D trace is shifted between the quiet hours preceding and following the change. In figure 25, the level shift is apparent in the north component at Churchill as well as in D at College, Barrow, and Murmansk. In references 3 and 6, it is also obvious that the plasma spectra shifts to higher energies and greater flux over an interval centered on 2200 UT, March 26.

Although it is tempting to pursue explanations for the details of these correlations (e.g., treating the reading at 2201 UT in terms of a neutral sheet in a solar stream or cloud), the data is not sufficiently comprehensive. From a more general view, the correlations do suggest spatial structure in the solar wind and also that changes in this structure have a rather direct influence on magnetic activity at the surface of the earth, even during periods that will normally be called magnetically quiet. Thus, studies of small world-wide effects on surface magnetograms, such as the study of simultaneous changes by Nishida and Jacobs,[16] may yield an index for the variability of the solar wind when these studies are extended to total intervals of simultaneous, similar behavior.

Sudden Commencement Correlations

As noted on Page 921, a magnetic storm sudden commencement (SC) occurred at the earth's surface at 1503 UT, March 27 approximately 29 hours after a Class 3 flare near the east limb of the sun. On the basis of the change in plasma energy spectra and total flux[3,6] and the rise in field intensity after 1503 UT (figure 13), it is logical to associate the increased field intensity with the SC at the earth's surface. Identification of an exact origin time at the satellite is not possible as the field increased gradually from 1503 to 1510 UT before making a larger change between 1510 and 1512 UT. Thus, the origin time can only be specified as 0 to 7 minutes later at the satellite than at the earth's surface with the largest single change occurring 7 to 9 minutes later. The transit

time for the Class 3 flare gives an average straight line propagation velocity of 1460 km/sec, equivlent to 11.1 kev protons. The same radial velocity would give a 2.3 minute difference in arrival time at the earth and satellite due only to their different distances from the sun. Approximately one-half this velocity, which agrees with the plasma energies measured after the SC,[3,6] gives a time difference of about 5 minutes. Thus, the assumption that the origin time at the satellite is close to 1503 UT, and not 1510 UT, gives an approximate *upper limit* of 5 minutes for propagation of the SC from outside the earth's field to the earth's surface.

Figure 27 illustrates the SC appearance. At low latitudes the SC was distinct, but the main phase of the storm was weak. At College, Alaska, the SC produced an abrupt change of about 300 gammas in the horizontal component which was accompanied by the onset of strong absorption on riometer records (Leinbach, personal communication) similar to other SC events in which radiation has been simultaneously detected at balloon altitudes.[17] The fact that the SC is not apparent at Pt. Barrow and is small at Sitka suggests that the initial impulse was localized. An abrupt change, 280 gammas in H, also occurs at Godhavn, Greenland. The initial effects at other stations examined are in general much smaller than those at College, Godhavn, and Thule. This suggests that solar particles with energies greater than the measured proton energies[3,6] may accompany the plasma cloud (or stream) and that these reach particular high latitude regions by a rather direct path. This could also be suggested on grounds that the SC effect at the satellite represents an increase of only a factor of 2 in the observable plasma flux relative to a number of other periods during the flight.

It also presents the question as to whether or not the magnetic fields observed by Explorer X are not also affected by particles which have energies outside the range of detection of the plasma probe. This question arises when the data prior to the SC is examined. From figure 13 it is apparent that the direction of the field vector was not significantly changed during the initial rise in field intensity following the SC time. Instead the field orientation up to half an hour after the SC appears to follow a slow shift (particularly

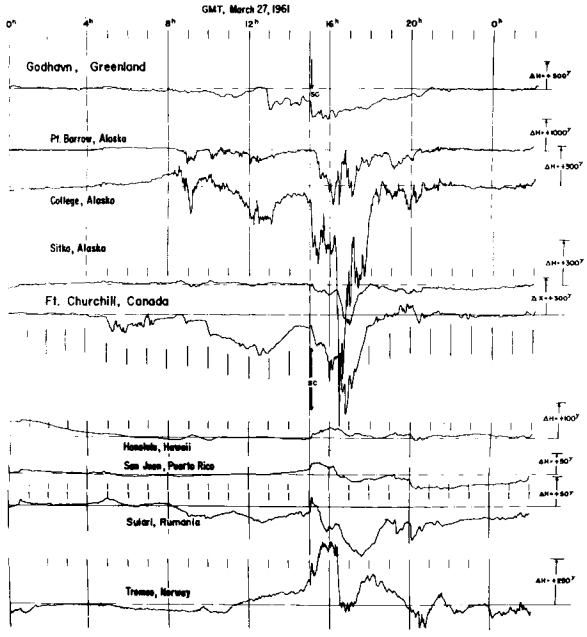


FIGURE 27.—Magnetograms illustrating the sudden commencement at 1503 U.T., March 27 (note scale differences).

apparent in the angle ϕ) that started two hours earlier. This shift was temporarily broken between 1435 and 1450 UT by a large change in field orientation that is not accompanied by either a magnitude change or a change in plasma flux. The significance of this change is obscure. However, during the preceding time (figure 13), it is apparent that the field near 1300 UT made a rapid transition to the general orientation it had during and following the SC. The change appears to start between 1257 and 1259 UT with a shift most evident in the θ angle. As shown in figure 27, the horizontal field intensity at Godhavn, Greenland, geomagnetic latitude 79.8 degrees, drops 360 gammas between 1250 and 1300 UT (note the sensitivity scales to the right of figure 27). Thus, the time difference between the initial changes at Godhavn and the satellite is about 7 minutes and the time difference between minimum H at Godhavn, about 1300 UT, and minimum intensity at the satellite, 4.2 gammas at 1301 UT, is less than 2 minutes. These time differences are similar to those noted at the time of the SC. If we assume that the change at the satellite and at Godhavn is not just a meaningless coincidence,

there is a basis for arguing that the continuity of field orientation before and after the SC is a consequence of particles with energy significantly greater than 2.3 kev in advance of the main plasma cloud. The energy density of the higher energy particles (e.g., in the still unmeasured range of 5 to 50 kev) would have to be greater than the field energy density to produce the magnetic effect at the satellite. At still higher energies, above 100 kev, Explorer XII (References 18 and 19) has in at least one instance shown the existence of higher energy particles up to 30 minutes in advance of an SC but with insufficient energy density to cause a large field change at the satellite.

Although there is little documented statistical evidence for large field changes at the earth's surface prior to SC's, it should be noted that these changes may be localized as in the case above. Godhavn, in this case, is near the 10:40 local magnetic time meridian which could be significant on the basis of the following considerations: (1) the proximity to the expected location for a Chapman-Ferraro "horn"[20,21,22] for particles incident along the sun-earth line or from a direction to the west of the sun as seen from noon on the earth; (2) the promixity to the 10:00 meridian which Wilson and Sugiura[23] have found to be a dividing plane between opposing rotations of the magnetic vector at times of sudden commencements; and (3) the coincidence between the plane of the Godhavn meridian and the average orientation of auroral arcs over the polar cap.[24]

The possible importance of particles with energy greater than 2.3 kev can also be argued on the basis that the observed plasma flux[3,6] between 1300 and 1503 UT was lower than average and also variable which suggests that it was not dominant in controlling the relatively stable field over most of the interval. These arguments for high latitude effects at the earth's surface preceding the SC and the possible importance of particles with energy significantly greater than 2.3 kev are not conclusive. They do suggest that it would be instructive not only to make additional measurements but also to systematically study magnetograms for events prior to SC's. The measurements following the SC show that the field intensity (figure 13) and plasma flux remain at a higher level for at least 4 hours, after which battery power was lost. The field direction, summarized in figure 19c, was, in general, such that the vector pointed to the west of the sun and varied irregularly, but somewhat systematically, above and below the ecliptic plane.

THE GEOMAGNETIC CAVITY

Cavity Models

The problem of interaction between the geomagnetic field and an ionized but neutral corpuscular stream was formulated in 1930 by Chapman and Ferraro[20,25] to explain SC's and the initial phase of magnetic storms and has subsequently been treated by them in additional detail.[26,27] Recently, this problem has been approached by a number of investigators (references 21, 22, 28, 29, 30, 31, 32, and 33 among others) for cases of steady plasma flow and isotropic plasma pressure. Computations have been carried out for both two and three dimensional cases and in most cases it is assumed that the magnetic pressure just inside the cavity surface $B_s^2/8\pi$ is balanced by the plasma pressure just outside the cavity. The surface is in reality taken to be a thin current sheath having the necessary strength and geometry to identically cancel the geomagnetic field that would exist external to the same surface in the absence of a plasma pressure. Thus, the problem is formulated so that the geomagnetic field does not have a component normal to the surface at the boundary location. It is generally assumed that the plasma pressure inside the cavity surface is negligible and in nearly all computations the possible effects of an external field of solar-interplanetary origin are completely ignored.*

For a uniform solar stream with proton velocity v, mass m, and number density n, the equilibrium condition for the cavity surface is computed from

$$B_s^2/8\pi = 2mnv^2\cos^2\zeta, \qquad (1)$$

where ζ is the angle between the normal to the surface and the velocity vector of the stream prior to encountering the earth's field and B_s is

*The exception to this is Dungey's (reference 33) two-dimensional calculation for a selected orientation of an external interplanetary magnetic field.

the total magnetic field intensity at the boundary. Specular reflection of the incident plasma is assumed and although there is some concern over the factor 2 in the equation above (see references 22 and 29) it has been used by most investigators.

The computation of the surface geometry then becomes one of finding numerical solutions to the boundary value problem subject to additional approximations. The approximation

$$B_s = 2fB_t, \qquad (2)$$

where B_t is the component of the earth's dipole tangential to the surface, is often used to specify the condition that the field inside the cavity cannot have a component normal to the surface at the boundary. For arguments regarding the value of f, usually taken to be unity or less, and the numerical techniques and approximations then employed, the reader should consult the references listed, and especially the work by Beard[29] and Spreiter and Briggs.[22]

Cavity Geometry Relative to Cavity Models

In the preceding sections arguments have been advanced that the Region A magnetic fields can best be interpreted as fields within the geomagnetic cavity. The arguments are based almost entirely on: (1) the similarity between the Region A field direction and the field direction observed between 12 and $22R_e$ (page 926); (2) the deduction that the Region A field intensities decrease with distance from the earth when importance is attached to the earth's daily rotation (pages 927 and 928); and (3) the absence of a detectable plasma flux when the field has the Region A orientation. If this interpretation is correct, the theoretical models can be compared with the measurements to check the degree of agreement.

Considering first the geometry of the cavity, it is apparent that the measurements support the theoretical prediction that the solar wind extends the geomagnetic field to great distances on the dark side of the earth. If, however, we then assume that the plasma velocity vector is directed exactly along the sun-satellite line, the dimensional agreement between theory and measurement becomes increasingly poorer with distance. This is illustrated in figure 28 using Spreiter and

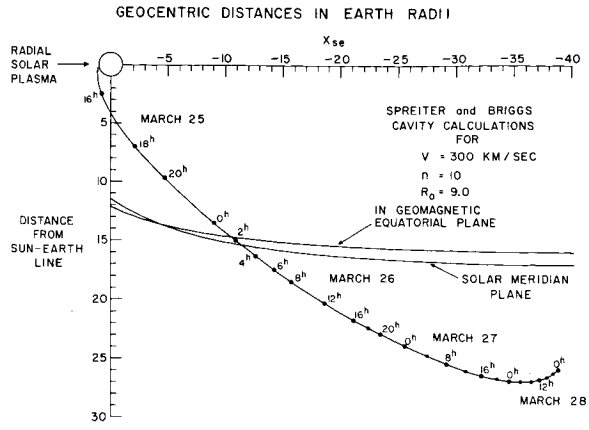


FIGURE 28.—Cavity dimensions given by Spreiter and Briggs relative to trajectory dimensions.

Briggs's[22] computations of the cavity dimensions in equatorial and meridian planes for a typical observed plasma flux.[3,6*] The fact that Region A fields are observed in intervals between 0600 and 1300 UT, March 27, as well as between 0600 and 1800 UT, March 26 (figures 20 and 21), obviously requires, under the assumptions of radial plasma flow, that a symmetrical cavity be roughly conical rather than cylindrical at great distances along the trajectory. Also, rotating the position of the earth's dipole axis relative to a radial stream in the Spreiter and Briggs model does not appreciably alter the cavity dimensions shown in figure 28. This further implies that the model calculations do not provide an explanation for the behavior of field intensities in Region A, which were discussed earlier. In general, from these and other arguments, it is apparent that dimensional agreement between the measurements and the models cannot be reached using the conditions that the plasma is directed along the sun-earth line and that the cavity formed is continuously symmetric relative to the earth-sun line. Assuming that the computations are basically correct for the assumptions made, there are a number of reasons why agreement might not be expected. The neglect of an external solar-interplanetary field in the model calculations is perhaps the most obvious, but until model calculations have been made either for the general case or a variety of selected cases the

*Beard's solution (reference 29) gives substantially the same dimensions in the region of interest here.

consequences of this neglect cannot be estimated with confidence. Second, although the plasma flux measurements are consistent with a flow directed in the sun-earth direction, flow from a direction to the west of the sun is not excluded.[3,6] Third, the measurements do not exclude the possibility of plasma outside the observable energy range arriving from either the radial or other directions. It may, in fact, be essential to postulate a variable component of the solar stream arriving from a direction to the west of the sun to explain why the cavity moves back and forth across the satellite trajectory between 0600 and 1300 UT, March 27, and 0600 and 1800 UT, March 26. An alternative explanation for the intermittent reappearance of Region A is that the degree of symmetric compression of the cavity differs between times when the satellite is inside or outside the cavity in response to changes in the solar wind pressure. This explanation appears to conflict with the fact that Region A fields are not observed during the periods of minimum magnetic activity at the earth's surface when we might expect the cavity to be the least compressed. This conflict may not exist, however, as the quiet magnetic periods occur within the 12-hour period in which Region A fields never appear. This may be attributed to a diurnal effect (see, e. g., page 927). Similarly, the conflict may not exist on the grounds that Region A fields do not appear after the SC.

These arguments do not lead to a consistent picture. They do, however, make it evident that relative to existing cavity models one must either postulate a non-radial component in the solar stream or postulate that the external field has a large effect on the cavity geometry. The alternative is to assume that the cavity calculations are wrong in not predicting a broad conical cavity back to the earth, even in the idealized formulation.

Cavity Boundary Conditions

Interpreting the Region A magnetic fields as fields within the cavity, we can compare the boundary conditions with those usually assumed in theory and was outlined on page 933. In doing this, it is immediately obvious that the condition $B_s = 2fB_t$ (equation 2) is not observed if we take $f \leq 1$ and compute B_t for a dipole field. Instead, an $f > 3$ is required when B_t for a dipole is assumed. On the other hand, if we use typically observed values for B_s, n, and v in equation 1, and compute ζ, assuming a radial $\bar{v}$, the range of ζ is roughly 55 to 85 degrees. This agrees approximately with the range of 50 to 70 degrees that would typically represent the angle between the sun-earth direction and the normal to a roughly conical surface containing the trajectory. The agreement would be improved if the plasma was assumed to arrive from a direction slightly to the west of the sun. The question of agreement also remains open in that equation 1 assumes specular reflection whereas the plasma measurements[3,6] have not revealed a reflected component. This could be a geometrical circumstance in which the incident and reflected components superimpose and are not separable. This seems unlikely unless the cavity geometry is the same each time it crosses the trajectory and is also very smooth. Alternatively, if ζ in equation 1 is always large (i.e., incidence from a direction west of the sun), the incident and reflected components would not be resolved. More exact tests than the above require continuous, rather than time spaced, measurements at the time of cavity crossings.

The limitation of noncontinuous data becomes even more restrictive in checking the model assumption that the field at, but inside, the boundary cannot have a component normal to the surface. Bridge, Bonetti, et al.[6] concluded that the field must have a normal component on the grounds that the normal to the surface, defined by the cross-product between two successive measurements at the time of a cavity crossing, was not compatible with a surface that would permit detection of radial plasma flow. However, widely different results are obtained, depending on the exact choice of measurements and this makes the test questionable. We may similarly look at the cross-product of average vectors in Region A relative to Regions B and C, but this involves the assumption that the field external to the cavity is perfectly wrapped about the cavity surface without maintaining even a fractional resemblance to its original geometry. The assumption becomes untenable if the average distances from the cavity surface are large as might be expected from the

duration of most of the intervals in which the field is in one of the three regions. In addition to these difficulties in using a cross-product, it involves in interpretation the assumption that the cavity has a rather simple symmetry which, though convenient for illustration pruposes, may not be justified in view of the apparent diurnal behavior (p. 928) and various other features already discussed. As an illustrative example, the projection of Region B vectors in the X_{se}, Y_{se} plane of figure 22 suggests that the field extended almost normal to a symmetric conical cavity; but if viewed from the sun ($Y_{se}Z_{se}$ plane), it is apparent that only a slight distortion of such a cavity would place the vectors in front (i.e., on the sunward side) of the assumed cavity.

Other arguments can be used that suggest that the field inside the cavity but near its surface is tangential to the surface. The most obvious is the fact that the Region A vectors (figure 21) show only small changes in orientation from 22 to $41R_e$, and these change appear to have a diurnal dependence. Another argument against having both a symmetric cavity and geomagnetic field lines crossing the cavity surface, is that it becomes exceedlingly difficult to account for the total magnetic field flux. To explain the observed field intensities in the case of a symmetric cavity, when no lines are allowed to cross the surface, it is necessary to assume that all lines of force intersecting the earth at latitudes of 72 degrees or higher extend through a plane perpendicular to the earth-sun line approximately $30R_e$ behind the earth within the confines of the conical cavity. That is, only an insignificant fraction of the flux can pass through the equatorial plane between 10 and 30 or more R_e. Thus, unless the flux is highly concentrated near the cavity surface and not distributed throughout the volume back of the earth, it is not tenable to postulate both a symmetrical cavity and lines of force crossing the surface. It also seems unlikely that there would be two highly preferred directions, Regions B and C, for the field orientation outside the cavity if these fields were to be attributed to geomagnetic lines crossing the boundary. Similarly, if a significant fraction of the field intensity outside the surface was due to field sources inside the surface, we would expect to find a decrease in intensity in Regions B and C with distance, and this is not the case.

In summary, there is little reason to believe that the geomagnetic lines cross the surface and, indeed, observations indicate that they do not. Thus, it is not necessary to believe that the measurements contradict the common assumption that the geomagnetic lines do not extend into the highly conductive solar plasma. This does not preclude the existence of a sheath at the boundary in which mixing and diffusion occur as this must certainly take place within a finite volume. It suggests that the sheath thickness is small compared to the average distance from the boundary when the satellite is in one of the characteristic field configurations, Regions A, B, and C (figures 19 and 20). As the measurements do not reveal either the distances from the boundary or the boundary thickness, we can only estimate that on the dimensional scale involved the average distances from the cavity could easily be several earth radii and at least several thousand kilometers, and thus very large compared to typical values computed for the boundary thickness, even for cases of surface instability. (See, for example, reference 34.)

FIELDS EXTERNAL TO THE GEOMAGNETIC CAVITY

In an earlier section the Region B and C fields were interpreted as fields of solar-interplanetary origin whose orientations may be significantly affected by the proximity to the geomagnetic cavity. The degree of cavity influence was taken to be uncertain and could be negligible, especially between 1830 UT, March 26, and 0530 UT, March 27. Continuation of this interpretation of the measurements should reveal some of the properties of the solar wind and the behavior of the associated fields.

It is apparent that the field energy density is typically a factor of 5 to 10 less than the observed plasma energy density and thus it can be assumed that the field is carried by the solar wind. The measurements also suggest that the solar wind has both small and large scale variability. The small scale variability is suggested by the changes that take place between successive samples at 2-minute intervals. The large scale variability is

suggested by the tendency for a given field configuration to persist for fractions of an hour to several hours and appears both as a movement of the cavity across the trajectory, typically transitions between Region A and Region C configurations, and as disruptions of the Region B configuration during the 12-hour period when plasma disappearance and Region A fields are not observed. The deviations from the Region B configuration, discussed on pages 193 and 194, are particularly important in interpreting the large scale structure in that each change away from the Region B condition is accompanied by a shift to a higher energy plasma spectrum. This presents the possibility, when extrapolated to very quiet conditions, that a very weak solar wind is accompanied by a spiral field which at 1 AU lies roughly in the ecliptic plane at nearly right angles to the solar radial direction as predicted for this simple case by Parker[35] and others. The data suggests, however, that such a stable configuration may be quite rare and that to retain Parker's picture it would be closer to reality to view the interplanetary structure along the lines of a series of small "blast waves" from the sun (Parker[36]). The interplanetary media would then be broken into annular regions moving outward from the sun and a typical annular width from the Explorer X data would be of the order 0.01 AU from $v(\Delta t)$, where Δt is the time interval for the region to cross the satellite and v is the observed plasma velocity. At the other extreme, the structure could be entirely filamentary with the individual filaments oriented roughly at the stream angle $\tan^{-1}(\omega r/v)$ with respect to the solar radial direction, where ω and r are, respectively, the sun's angular velocity and the distance from the sun. Filament dimensions would typically be $\omega r(\Delta t) \times \cos[\tan^{-1}(\omega r/v)]$ in this case, and thus not greatly different than 0.01 AU. With filaments of this small scale, agreement between filament orientation and field direction can hardly be expected as the range of observed velocities would certainty lead to intersecting flow between adjacent filaments in the space between the sun and the earth. Similarly, in a "blast wave" picture, the flow would necessarily involve higher velocity regions overtaking the lower velocity regions and the resulting field configuration can hardly be expected to fit any simple model. Thus, the data is obviously not definitive in selecting between "blast wave" or filamentary structures and similarly does not distinguish between these structures and others that might be visualized from a "magnetic bottle" (e.g., reference 37) picture. The picture that does emerge is one of considerable structure in the solar wind under average conditions of solar activity. The correlations with surface magnetic activity also lead to this suggestion. The next question is whether or not these structures are related to minor solar events on the sun. As illustrated in figure 14, the time intervals between Class 1 flares prior to the flight have some similarity to the observed duration of a given field configuration. Correlations cannot be justified, but the coincidence suggests that it might be worthwhile to attempt statistical correlations between numbers and locations of minor solar flares and occurrences of simultaneous world-wide magnetic field changes of the type discussed on pages 192 and 193. A correlation of this type would not be found by the usual techniques, as it requires detailed study of a large number of magnetograms from observatories widely distributed in longitude.

It may not be sufficient to look for events on the sun to explain the solar wind structure. Instead dimensions of the order 0.01 AU at distances of 1 AU may be related to a more fundamental and continuous property of the sun and solar corona. Some suggestion of this comes from McCracken's[38] deduction that scattering centers with dimensions of 0.01 AU could account for observed angular distributions in cosmic ray flare effects. Dimensions of 0.01 AU, or less, at distances of 1 AU are also indicated by extrapolating recent observations[39] of the scattering of radio emissions passing through the solar corona at distances out to 0.5 AU from the sun. These observations[39] are particularly suggestive in that a close correlation was also found between detection of scattering at great distances in equatorial regions (0.25 to 0.5 AU) and the areas and intensities of calcium plages. Thus, if the basic scattering structure is that of filaments which are extensions of visible solar rays and plage activity leads to greater density contrast between filaments there is a reasonable basis for suggesting that this is the type of structure observed by Explorer X.

From figures 13 and 19, it is apparent that the

solar stream from the Class 3 flare effectively displaced the existing field structure and the average field orientation (figure 19) became more oriented in the ecliptic plane with time. The rather large angular deviations show, however, that the field within the stream was by no means higher regular and as was discussed on pages 196 through 198, we cannot assume that the SC represents a sharp stream boundary. Depending on the choice of origin of the stream, 1300 or 1503 UT (see pages 930 through 931), a minimum stream dimension corresponding to 4 to 6 hours in passage can be stated, but the maximum remains uncertain due to the loss of battery power.

It is also apparent that the ratio of field energy density to observed plasma energy density after the SC is essentially the same as observed during other periods when plasma was present between 22 and $40R_e$. Thus, if it is assumed that the proximity of the geomagnetic cavity does not affect the measurements after the SC, this may indicate that the Region B and C fields are not as affected by the cavity proximity as might be supposed on other grounds.

Comparison of the Region B and C fields with the measurements of Pioneer V (1960α) does not show appreciable agreement in detail.[40,41,42] Field directions are not readily compared in that Pioneer V measured only the magnitude perpendicular to the spin axis and thus fields perpendicular to the ecliptic could not be distinguished from fields in the ecliptic plane normal to the radial solar direction.[43] However, it is apparent in comparing magnitudes that the Explorer X components in the same plane would in general give values in excess of the 2.7 to 5.0γ fields typically observed with Pioneer V. Similarly, Pioneer V apparently did not see evidence for large scale structure, except at times of magnetic storms, although the data intervals may in general have been too short for this to be apparent. Comparison of magnetic activity indices during the two flights also makes it clear that the Explorer X measurements prior to the SC occurred when the indices were comparable to those of the quiet days during the Pioneer V flight. Thus, differences are not readily explained in terms of the level of magnetic activity. Resolution of these differences can be attempted by assuming that the Region B and C fields show greater magnitude as a consequence of crowding of the field lines around the geomagnetic cavity; however, various features argue against this and the differences will best be resolved by future measurements.

ALTERNATIVE INTERPRETATIONS

At various points in the preceding discussion, alternative interpretations were presented, usually with respect to a particular problem, but no attempt was made to consider all possibilities. There are alternatives which would lead to grossly different interpretations that should be mentioned even though space will not permit extensive discussion. These are noted below:

1. If major differences in the plasma energy spectra existed between adjacent filaments in the solar wind, it is possible that the Region A field intervals could be solar wind filaments in which particle energies are greater than 5 kev. In this case, a cavity picture more consistent with theoretical cavity models could be obtained from the standpoint that a cylindrical surface, crossed once at 0530 UT, March 26, would be adequate and major diurnal distortions would not be indicated by the measurements. The principal difficulty is obviously the lack of observational evidence for higher energy plasma.

2. As in (1) above, in suggesting large velocity differences between adjacent filaments but instead treating the plasma-free regions as cool filaments, Parker (personal communication) has suggested that hot (solar wind observed) filaments and cool (solar wind not observed) filaments could be developed in adjacent columns low in the solar corona and continued outward to great distances with the hot filaments effectively drawing the fields outward in the cool filaments. Extending this further, we might suggest that the cool filaments superimpose with the earth's field to explain the similar orientation of Region A fields and the field between 12 and $22R_e$. The principal difficulty in explaining the observations is perhaps accounting for the regularity of the Region A fields.

3. As an alternative to explaining the transitions between Region A and Region C fields in terms of either movement of the entire cavity or differences in compression of a symmetrical

cavity, it can be suggested that the cavity surface is highly unstable and irregular or that the surface exhibits long period oscillations. As noted on page 920, and previously illustrated,[1,2] the measurements between 0530 and 1330 UT, March 26, typically show short period field fluctuations when plasma is observed that might be interpreted as an unstable surface behavior. Uncertainty arises in interpreting this as a typical behavior in that the same time interval coincides with an extended period of field fluctuations at the earth's surface as discussed on page 193.

4. W. I. Axford (personal communication) and Kellog,[44] following the approach of Zhigulev,[26] have proposed that the observed field changes can be explained by the existence of a collisionless shock front spreading outward in the form of a bow wave from a vertex position located several or more earth radii in front of the magnetosphere at the subsolar point. The existence of the shock considerably reduces the plasma bulk velocity, perhaps to random motion (W. I. Axford, personal communication), but the motion presumably again becomes radial at large distances behind the shock. The magnetic field behind the shock and outside the magnetosphere is presumably of solar-interplanetary origin but greatly modified by the presence of the shock. Comparison in any detail with the measurements is currently impossible as the postulated conditions have not been adequately developed from either an observations or theoretical approach.

5. As discussed on pages 922 through 925, the data does not suggest the existence of a symmetrical ring current at distances greater than $4R_e$. This does not preclude the existence of current systems within the cavity such as a "sheet" configuration in the equatorial plane extending over many earth radii. By careful selection of the distribution of current density and closing the circuit on the nightside of the earth, probably along field lines to high latitudes, a fit to the observations could probably be obtained. Explanation along these lines might also explain the outward spreading of the cavity back of the earth, the total flux required in the cavity, and other features which are difficult to explain in a model involving no external magnetic field, radial plasma flow, and a conically symmetric cavity. The rather highly selective current distribution required to fit the observations is the principal difficulty in seeking explanation along these lines.

DISCUSSION

The measurements taken in radiation belt regions are not inconsistent with a weak diamagnetic ring current below the magnetic shell intersecting the equator at $3.8R_e$. Questions regarding the possible existence of meridional currents and the extension of the cavity field suggest that simple interpretations may not be adequate.

The measurements between 8 and $22R_e$ and during periods when plasma was not observed between 22 and $41R_e$ can be interpreted in terms of a large-scale cavity field that decreases in intensity with distance as $F(r_1)(r_1/r)^{\chi}$ with $0.6<\chi<1.0$ for $r_1=11.5R_e$, $r>11.5R_e$, when it is assumed that the intensity observed depends on the position of the dipole axis relative to the solar stream.

At distances greater than $22R_e$, and prior to the SC, the field is typically oriented in one of three directions designated Regions A, B, and C. A distinctly different level of average field intensity is associated with each of the three orientations. The Region A fields correlate with the absence of detectable plasma and are interpreted as fields within the cavity. The Region B and C fields are observed only when plasma is observed. The occurrence of Region B fields coincides with both a different degree of magnetic activity on the earth's surface and a different position of the dipole relative to the solar stream than the A or C fields. This double correlation complicates interpretation as the relative importances are not separable.

World-wide changes in the magnetic field at the earth's surface characterized by small increases in the horizontal component are found to correlate with field changes at the satellite.

A change in field intensity at the satellite occurring 0 to 7 minutes after a SC at the earth's surface is identified as a SC effect. However, the direction of the field vector up to half an hour after the SC followed a slow shift in orientation that began abruptly 2 hours before the SC in coincidence with an abrupt change in the magnetic

field at high latitudes at the earth's surface. This feature and others present questions regarding the possible importance of particles with energies greater than those measured suggest that these particles may arrive in advance of the lower energy plasma.

The cavity characteristics indicated by the measurements do not agree in detail with theoretical cavity models. Agreement might be obtained if it were assumed that the average solar stream was incident from a direction slightly to the west of the sun and above the ecliptic plane or if there was an intermittent component of the flow from this direction. Alternatively, agreement should probably not be expected as the theoretical models do not take into account the existence of a magnetic field outside the cavity.

On various grounds the fields measured external to the geomagnetic cavity are interpreted as fields of solar-interplanetary origin. Uncertainties arise in estimating the extent to which these fields are influenced by the proximity to the cavity.

The tendency for a given field condition to persist for fractions of an hour to several hours, which appears both as a movement of the cavity across the trajectory (typically transitions between Region A and Region C fields) and as disruptions of the Region B fields during a 12-hour period when plasma is always present, is interpreted in terms of structure in the solar wind. Dimensions of this structure would typically be of the order 0.01 AU. Arguments tend to favor the view that these structures may be filamentary extensions of coronal rays, but this is only one of several possibilities.

In addition to the points just cited, alternatives in interpretation have been noted to emphasize the lack of uniqueness that ensues from the limited state of knowledge of the interactions of plasma with weak magnetic fields, and from the restrictions of measurement along a single trajectory.

REFERENCES

1. Heppner, J. P., Ness, N. F., et al., "Magnetic Field Measurements with the Explorer X Satellite," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961; II. *Main Sessions,* Tokyo: Physical Society of Japan, 1962, pp. 546–552.
2. Ness, N. F., Skillman, T. L., et al., "Magnetic Field Fluctuation on the Earth and in Space," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961; II, *Main Sessions,* Tokyo: Physical Society of Japan, 1962, pp. 27–33.
3. Bridge, H. S., Dillworth, C., et al., "Direct Observations of the Interplanetary Plasma," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961; II, *Main Sessions,* Tokyo: Physical Society of Japan, 1962, pp. 553–560.
4. Finch, H. F., and Leaton, B. R., "The Earth's Main Magnetic Field—Epoch 1955," *Monthly Not. Roy. Astronom. Soc., Geophys. Suppl.,* **7**(6): 314–317, November 1957.
5. Cain, J. C., Shapiro, I. R., et al., "Measurements of the Geomagnetic Field by the Vanguard III Satellite," NASA Technical Note D–1418, 1962.
6. Bridge, H. S., Bonetti, A., et al., "Explorer X Plasma Measurements," Paper presented at the 3rd Internat. Space Sci. Sympos., Washington, May 1962.
7. Dolginov, S. Sh., and Pushkov, N. V., "On Some Results of Magnetic Field Investigations in Space," Paper presented at the 3rd Internat. Space Sci. Sympos., Washington, May 1962.
8. ———, and Eroshenko, E. G., et al., "Magnetic Measurements with the Second Cosmic Rocket," in: *Artificial Earth Satellites,* **5,** 490–502, New York: Plenum Press, 1961.
9. Heppner, J. P., Stolarik, J. D., et al., "Project Vanguard Magnetic Field Instrumentation and Measurements," in: "Space Research," *Proc. 1st International Space Science Symposium, Nice, January* 1960; Amsterdam: North-Holland Publ. Co., 1960, 982–99.
10. Cain, J. C., Stolarik, J. D., et al., "Vanguard III Magnetic-Storm Measurements," Paper presented at 43rd annual meeting, *Amer. Geophys. Union,* April 1962.
11. Cahill, L. R., "Study of the Outer Geomagnetic Field: Explorer XII," Paper presented at the 3rd Internat. Space Sci. Sympos., Washington, May 1962.
12. Davis, L. R., "Low-Energy Trapped Protons," Paper presented at 43rd annual meeting *Amer. Geophys. Union,* April 1962.
13. Davis, L. R., "Low-Energy Trapped Protons," Paper presented at 43rd annual meeting *Amer. Geophys. Union,* April 1962.
14. Rossi, B., "Interplanetary Plasma," Paper presented at American Physical Society Meeting, New York, January 1962.
15. Smith, E. J., Coleman, P. J., et al., "Characteristics of the Extraterrestrial Current System: Explorer VI and Pioneer V," *J. Geophys. Res.,* **65**(6): 1858–1861, June 1960.
16. Nishida, A., and Jacobs, J. A., "World-Wide Changes in the Geomagnetic Field," *J. Geophys. Res.,* **67**(2): 525–540, February 1962.

17. Brown, R. R., Hartz, T. R., et al., "Large-Scale Electron Bombardment of the Atmosphere at the Sudden Commencement of a Geomagnetic Storm," *J. Geophys. Res.*, **66**(4): 1035–1051, April 1961.

18. Hoffman, R. A., Davis, L. R., and Williamson, J. M., "0.1 to 5 Mev Protons and 50 Kev Electrons at 12 Earth Radii during Sudden Commencement on September 30, 1961," Paper presented at 43rd annual meeting *Amer. Geophys. Union*, Washington, April 1962.

19. Bryant, D. A., Cline, T. L., et al., "Cosmic Ray Observation in Space," Paper presented at the 3rd Internat. Space Sci. Sympos., Washington, May 1962.

20. Chapman, S., and Bartels, J., "Theories of Magnetic Storms and Aurorae," in: *Geomagnetism*, Oxford: Clarendon Press, 1940.

21. Dungey, J. W., "The Steady State of Champam-Ferraro Problems in Two Dimensions," *J. Geophys. Res.*, **66**(4): 1043–1047, April 1961.

22. Spreiter, J. R., and Briggs, B. R., "Theoretical Determination of the Form of the Hollow Produced in the Solar Corpuscular Stream by Interaction with the Magnetic Dipole Field of the Earth," NASA Technical Report R-120, 1961. Also, Theoretical Determination of the Form of the Boundary of the Solar Corpuscular Stream Produced by Interaction with the Magnetic Dipole Field of the Earth," *J. Geophys. Res.*, **67**(1): 37–51, January 1962.

23. Wilson, C. R., and Sugiura, M., "Hydromagnetic Interpretation of Sudden Commencements of Magnetic Storms," *J. Geophys. Res.*, **66**(12): 4097–4111, December 1961.

24. Davis, T. N., "The Morphology of the Auroral Displays of 1957–1958. 2. Detail Analyses of Alaska Data and Analyses of High-Altitude Data," *J. Geophys. Res.*, **67**(1): 75–110, January 1962.

25. Chapman, S., and Ferraro, V. C. A., "A New Theory of Magnetic Storms. I. The Initial Phase," *J. Geophys. Res.*, **36**(2): 77–97, June 1931; *J. Geophys: Res.*, **36**(3): 171–186, September 1931; *J. Geophys. Res.*, **37**(2): 146–156, June 1932; *J. Geophys. Res.*, **37**(4): 421–429, December 1932. Also, "A New Theory of Magnetic Storms. II. The Main Phase," *J. Geophys. Res.*, 38(2): 79–96, June 1933.

26. Chapman, S., "Idealized Problems of Plasma Dynamics Relating to Geomagnetic Storms," *Rev. Modern Physics*, **32**(4): 919–933, October 1960.

27. Ferraro, V. C. A., "Theory of Sudden Commencements and of the First Phase of a Magnetic Storm," *Rev. Modern Physics*, **32**(4): 934–940, October 1960. Also, "On the Theory of the First Phase of a Geomagnetic Storm: A New Illustrative Calculation Based on an Idealised (Plane not Cylindrical) Model Field Distribution," *J. Geophys. Res.*, **57**(1): 15–49, March 1952. Also, "An Approximate Method of Estimating the Size and Shape of the Stationary Hollow Carved Out in a Neutral Ionized Stream of Corpuscles Impinging on the Geomagnetic Field," *J. Geophys. Res.*, **65**(12): 3951–3953, December 1960.

28. Zhigulev, V. N., "O iavlenii magnitnogo 'otzhatiia' potoka provodiashchei sredy," *Akademiia Nauk SSSR. Doklady*, **126**, 521–523, 1959; and references cited therein.

29. Beard, D. B., "The Interaction of the Terrestrial Magnetic Field with the Solar Corpuscular Radiation. 2. Second-Order Approximation," *J. Geophys. Res.*, **67**(2): 477–483, February 1962.

30. Hurley, J., "Interaction of a Streaming Plasma with the Magnetic Field of a Two-Dimensional Dipole," *Physics of Fluids*, **4**(7): 854–859, July 1961.

31. Slutz, R. J., "The Shape of the Geomagnetic Field Boundary under Uniform External Pressure," *J. Geophys. Res.*, **67**(2): 505–513, February 1962.

32. Midgley, J. E., and Davis, L., "Computation of the Bounding Surface of a Dipole Field in a Plasma by a Moment Technique," *J. Geophys. Res.*, **67**(2): 499–504, February 1962.

33. Dungey, J. W., "The Interplanetary Field and Auroral Theory," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961. II. *Main Sessions, Tokyo*: Physical Society of Japan, 1962, pp. 15–19.

34. Parker, E. N., "Interaction of the Solar Wind with the Geomagnetic Field," *Physics of Fluids*, **1**(3): 171–187, May–June 1958.

35. ———, "Dynamics of the Interplanetary Gas and Magnetic Fields," *Astrophys. J.*, **128**(3): 664–676, November 1958.

36. ———, "Sudden Expansion of the Corona Following a Large Solar Flare and the Attendant Magnetic Field and Cosmic-Ray Effects," *Astrophys. J.*, **133**(3): 1014–1033, May 1961.

37. Gold, T., "Plasma and Magnetic Fields in the Solar System," *J. Geophys. Res.*, **64**(11): 1665–1674, November 1959.

38. McCracken, K. G., "The Cosmic-Ray Flare Effect. 3. Deductions Regarding the Interplanetary Magnetic Field," *J. Geophys. Res.*, **67**(2): 447–458, February 1962.

39. Slee, O. B., "Observations of the Solar Corona out to 100 Solar Radii," *Monthly Not. Royal Astronom. Soc.*, **123**(3): 223–231, 1961.

40. Coleman, P. J., Davis, L., and Sonett, C. P., "Steady Component of the Interplanetary Magnetic Field: Pioneer V," *Phys. Rev. Letters*, **5**(2): 43–46, July 15, 1960.

41. Greenstadt, E. W., "Magnetic Storms in Interplanetary Space as Observed by Pioneer V," *Nature*, **191**(4786): 329–331, July 22, 1961.

42. Coleman, P. J., Sonett, C. P., and Davis, L., "On the Interplanetary Magnetic Storm: Pioneer V," *J. Geophys. Res.*, **66**(7): 2043–2046, July 1961.

43. Sonett, C. P., "Hyperwaves, Shock-Like Phenomena in the Outer Exosphere," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961. II. *Main Sessions*, Tokyo: Physical Society of Japan, 1962, pp. 528–531.

44. Kellog, P. J., "Flow of Plasma Around the Earth," *J. Geophys. Res.*, **67**(10): 3805–3811, September 1962.

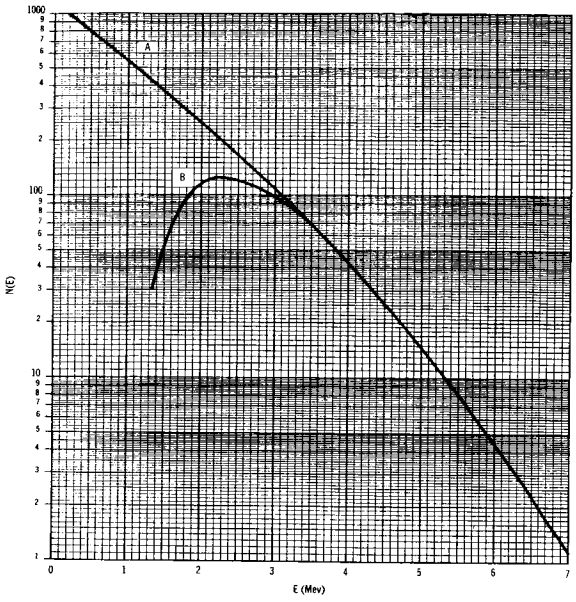


FIGURE 1.—Curve A is the fission energy spectrum and curve B the transmission energy spectrum for the Traac GM counter (0.66 gm/cm² wall).

Other detectors on Injun also give useful data sometimes, but often they are saturated and not usable. So far, little data have been analyzed from any Injun detectors except SpB.

Telstar I has on it a solid state p–n junction detector with pulse height analysis that selects electrons in different energy ranges from 0.2 to 1 Mev (private communication from W. Brown, Bell Telephone Laboratories). A lot of data have been reduced from Telstar I for two channels of the electron detector. This detector has given all the data currently available at high altitudes. It is directional, with an aperture half-angle of about 10 degrees. The fluxes are made omnidirectional by multiplying by the appropriate solid angle factor and then using a factor between 1 and 2 to correct, roughly, for the nonisotropic angular distribution.

Traac has a 302 GM counter shielded by 0.265 gm/cm² of Mg, which will count electrons of energy above 1.5 Mev (private communication from G. Pieper and L. Frank, Applied Physics Laboratory). It is essentially omnidirectional. Fluxes are obtained by dividing by $G_0=0.75$ cm² and correcting for saturation for high count rates.

ANALYSIS OF THE DATA

The data from these four satellites must be combined to form one overall picture of the artificial radiation belt. To do this assume that the energy spectrum of the electrons being counted is a fission spectrum. This is certainly the best guess. We will compare the data on this basis and see if there is agreement in the regions where direct comparison is possible. The fission energy spectrum $N(E)$ is shown in figure 1, curve A. A calibration of the Telstar I detectors at the Los Alamos Scientific Laboratory in a fission electron beam gives f, the fraction of fission electrons counted by the detectors, equal to 1/2.8 for the 240–340 kev channel and 1/6.0 for the 440–680 kev channel.

For Injun we have the experimentally determined factor $1/f$ of several thousand, by comparison of two detectors on board. The 213 GM counter has also been calibrated at Los Alamos with a fission electron spectrum (private communication from A. Petschek, H. Motz, and R. Taschek, Los Alamos Scientific Laboratory), and the factor f determined this way is 1/4000. We will use this factor in the present analysis. The Los Alamos tests show that the detector counts bremsstrahlung from electrons of several Mev rather than direct penetrating electrons. (If the shield had been carbon rather than lead, the counter would have counted direct penetrating electrons.)

For Traac, f is determined by considering the penetration of electrons through the detector shield of 0.265 gm/cm² of Mg and through the

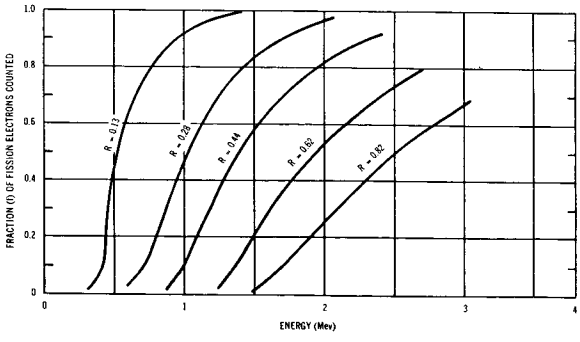


FIGURE 2.—The fraction of electrons of different energies that penetrate different shield thicknesses of Al.

THE ARTIFICIAL RADIATION BELT MADE ON JULY 9, 1962*

WILMOT N. HESS

Goddard Space Flight Center

The available information on the artificial radiation belt formed by the July 9, 1962, high altitude nuclear explosion is reviewed. Data from Injun (1961 $o2$), Telstar I (1962 $\alpha\epsilon 1$), Traac (1961 $\alpha o2$), and Ariel I (1962 $\eta 1$) are combined to form one picture of the artificial belt. The data are consistent to about a factor of 3. The flux map obtained in this way is used to calculate the flux encountered by several satellites. These show reasonable agreement with data on solar cell damage. Preliminary data on particle lifetimes are presented. Particles at $L>1.30$ are expected to last several years on the basis of coulomb scattering. Crude calculations of shielding are made to indicate the doses received inside various vehicles.

INTRODUCTION

On July 9, 1962, at 0900:09 UT a nuclear explosion of about 1.4 megatons was carried out at 400 kilometers above Johnston Island in the Pacific Ocean. This explosion produced, as was expected, an artificial radiation belt. However, the intensities in this radiation belt are considerably higher than were expected. Three days after the explosion the U.S.–U.K. joint satellite Ariel I (1962 $o1$) stopped transmitting. On August 2, Transit IV–B (1961 $\alpha\eta 1$) stopped transmitting: Traac (1961 $\alpha\eta 2$) stopped on August 14. Instruments on Ariel I, Traac, and Injun (1961 $o2$) showed large particle fluxes shortly after the explosion. It took about a month to start getting some grasp of the characteristics of the new radiation belt. This is a status report on the new belt as of September 12.

AVAILABLE DATA

The information that is available to form a picture of the new radiation belt comes mostly from particle detectors on the Ariel I, Injun, Traac, and Telstar I (1962 $\alpha\epsilon 1$) satellites. In addition to these data we can use the observed solar cell damage on satellites as an integral measurement of the trapped electron flux. Also, some data are available from dosimetry measurements.

Some of the original data about the enhanced trapped particle fluxes after the July 9 explosion came from the x-ray detector on the Ariel I satellite (private communication from A. Willmore, University College, London). This instrument was not designed to count charged particles and therefore its efficiency is uncertain. The data from it are quite useful in studying the time decay of the trapped flux and in locating contours of constant flux in B–L space.

Data received by the shielded 213 GM counter on Injun have been analyzed to give the first picture of the new radiation belt.[1] This counter is the background channel of the magnetic spectrometer, SpB. It has $3\frac{1}{2}$ gm/cm^2 of Pb shielding and about 1 gm/cm^2 of wall and miscellaneous shielding. It was supposed to give the penetrating background to be subtracted from the other channels of the spectrometer. This detector is now called on to provide quantitative information, and it has been calibrated after the fact. It is nearly omnidirectional. Fluxes are obtained from the count rates by dividing by $G_0=0.11$ cm^2.

*This is an abridged version of Technical Memorandum X-788, a confidential report entitled "The Artificial Radiation Belt."

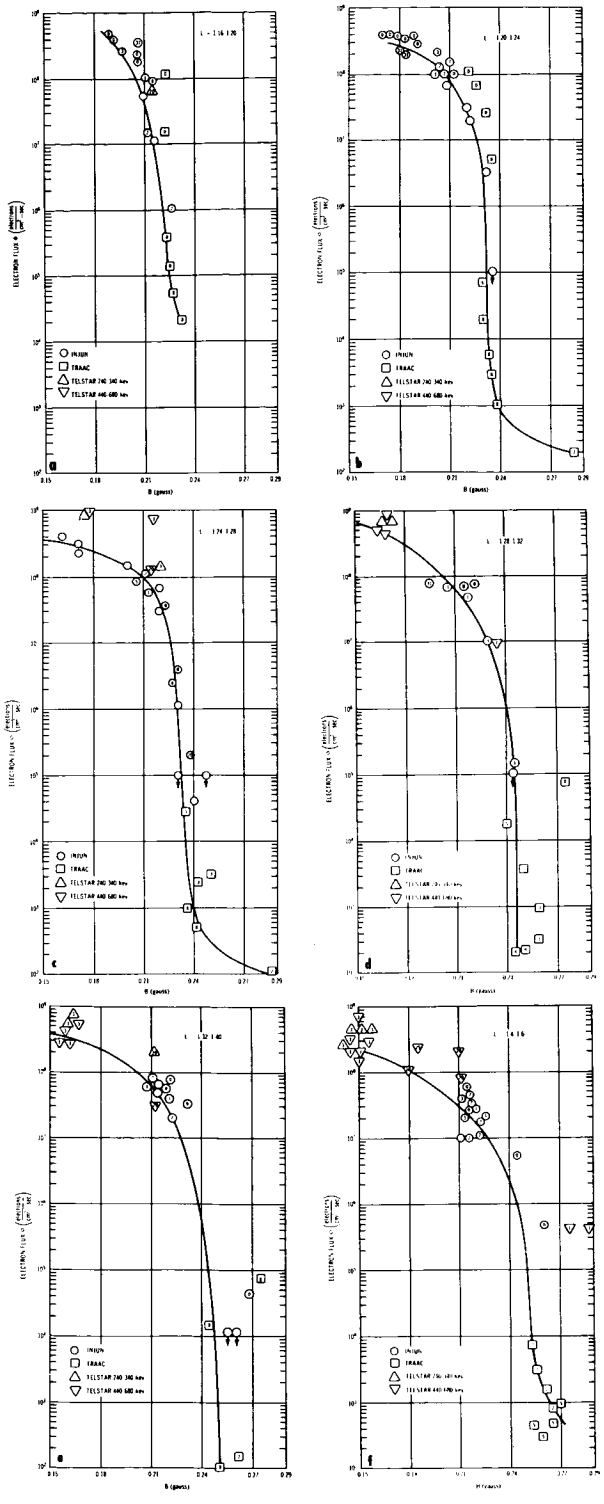


FIGURE 3.—The electron flux distributions along different field lines.

wall of 0.400 gm/cm² of stainless steel. Using the range straggling data[2] for Al we can get the fraction of electrons that penetrate a shield of given thickness, as shown in figure 2. The expression for the extrapolated range R is

$$R=0.526E-0.094.$$

This yields the absorber thickness that gives 10 percent transmission for electrons of energy E. For 50 percent transmission we multiply the energy by 1.38, and for 80 percent transmission increase the energy by a factor of 1.92. In this way we get the electron transmission spectrum, curve B in figure 1. The energies of the transmitted electrons are different from curve B, but the number transmitted is given correctly. The integral under this curve gives $f=1/5.5$ for the Traac counter. More information on shielding calculations is given in Appendix A.

Using the factors for the several detectors, we can calculate the total flux of fission electrons. In order to compare the different detectors, the total flux along several field lines (actually narrow ranges of L) has been plotted for different values of B (figure 3). These plots show that the different detectors agree fairly well in flux values. Avoiding the first day after the nuclear explosion (labeled by the number 0 inside the symbols on the graphs) we can see quite smooth trends in the data. The flux from Telstar I may be as much as twice as high as Injun fluxes. Traac and Injun agree quite well where comparisons are possible. In general, the data shows agreement to a factor of 2.

This agreement of the data shows two things: First, because the detectors give internally consistent results it seems likely that all the detectors are giving accurate information. Secondly, the assumption that the electrons have a fission energy spectrum appears to be correct. Of course it is possible that the energy spectrum is not a fission spectrum and also that the detectors are not in agreement, but it would have to be a peculiar combination of such effects that would give the agreement shown here. A comparison of the four channels of the Telstar I electron detector also indicates that the energy spectrum is fission-like up past 1 Mev.

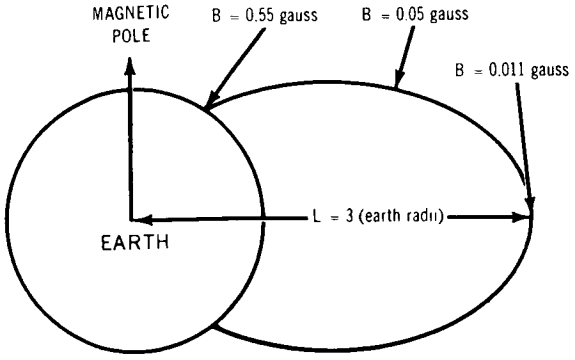


FIGURE 4.—The B–L magnetic coordinate system.

FLUX PLOTS

Now that it has been demonstrated that the energy spectrum is essentially a fission spectrum at least in the region of data overlap we can use all the counter data to construct a composite flux map in B–L space. As McIlwain has shown, these magnetic coordinates are the best way of organizing data about trapped particles.[3] L is constant along a field line in space and, for a dipole, is the distance from the center of the earth to the equatorial crossing of the line, in units of earth radii (figure 4). Values of L are calculated from the real values of the earth's field.

In constructing the flux map for $B>0.15$ gauss and for $L<2.0$ earth radii the graphs in figure 3 are used to locate the flux contours. The experimental data outside this B–L region are essentially all from Telstar I. There are several weeks data from Telstar I and considerable redundancy.

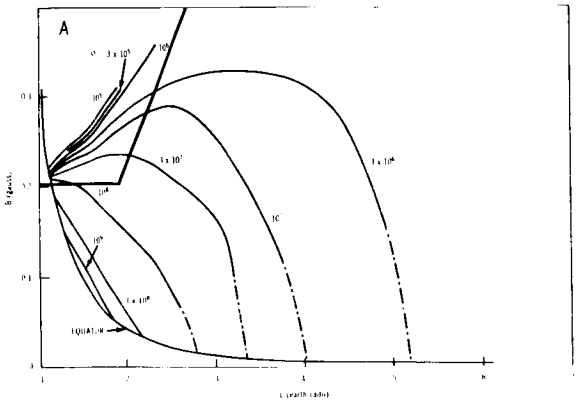


FIGURE 5.—The B–L map of electron fluxes.

The map made this way is quite complete. The data available in early September gave the flux map in figure 5. This map is for about 1 week after the explosion. There was considerably more flux at low altitudes at early times.

This same data plotted in R–λ coordinates, where

$$B=(M/R^3)[4-(3R/L)]^{\frac{1}{2}},$$

$$R=L\cos^2\lambda,$$

gives an equivalent dipole representation of the earth's field (figure 6). The maximum electron flux is about 2×10^9 elec/cm^2-sec. Integrating to get the total number of electrons stored in the field we find

$$\int\phi\,dV=2\times10^{26}\text{ electrons.}$$

About 40 percent of these electrons lie inside the 10^9 contour and about 60 percent of these electrons lie inside the 3×10^8 contour. It is not certain what fraction of these electrons are bomb-induced and what fraction are natural electrons. In this region around $L<1.5$ the energy spectrum seems softer than a fission spectrum.

The B–L flux map when plotted in terms of geographic coordinates gives the flux contours for different altitudes shown in figure 7.

VEHICLE-ENCOUNTERED FLUXES

A machine code has been developed which calculates the total number of electrons/cm^2 in the artificial radiation belt that strike a vehicle in space. This is done by calculating a point on the vehicle trajectory, transforming to B–L co-

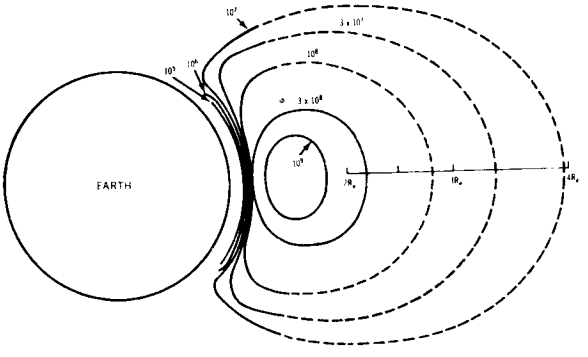


FIGURE 6.—The R–λ map of electron fluxes (an ideal dipole representation of the earth's field).

ordinates, looking up the electron flux, and integrating along the vehicle orbit. This has been performed and the encountered fluxes have been determined for all of the vehicles listed in Table 1. These fluxes have been transformed into r/day by using 3×10^7 electrons/cm^2=1r. The orbital elements of these vehicle trajectories are given in Table 2.

From the encountered fluxes in Table 1, we can learn several things. Let us first consider solar cell damage. The Bell Telephone Laboratories staff have studied this problem in considerable detail and prepared figure 8, which shows how different type cells are damaged by 1 Mev electron irradiation (reference 4 and private communication from the authors of that paper). Above about 0.5 Mev the electron damage is essentially independent of energy. Some care must be exercised in using this chart because of the variation in the characteristics of solar cells. We will assume all the electrons in the flux spectrum in figure 5 are greater than 0.5 Mev in estimating the solar cell damage.

About 20 percent degradation was needed by the blue sensitive p-on-n type cells on Ariel I to

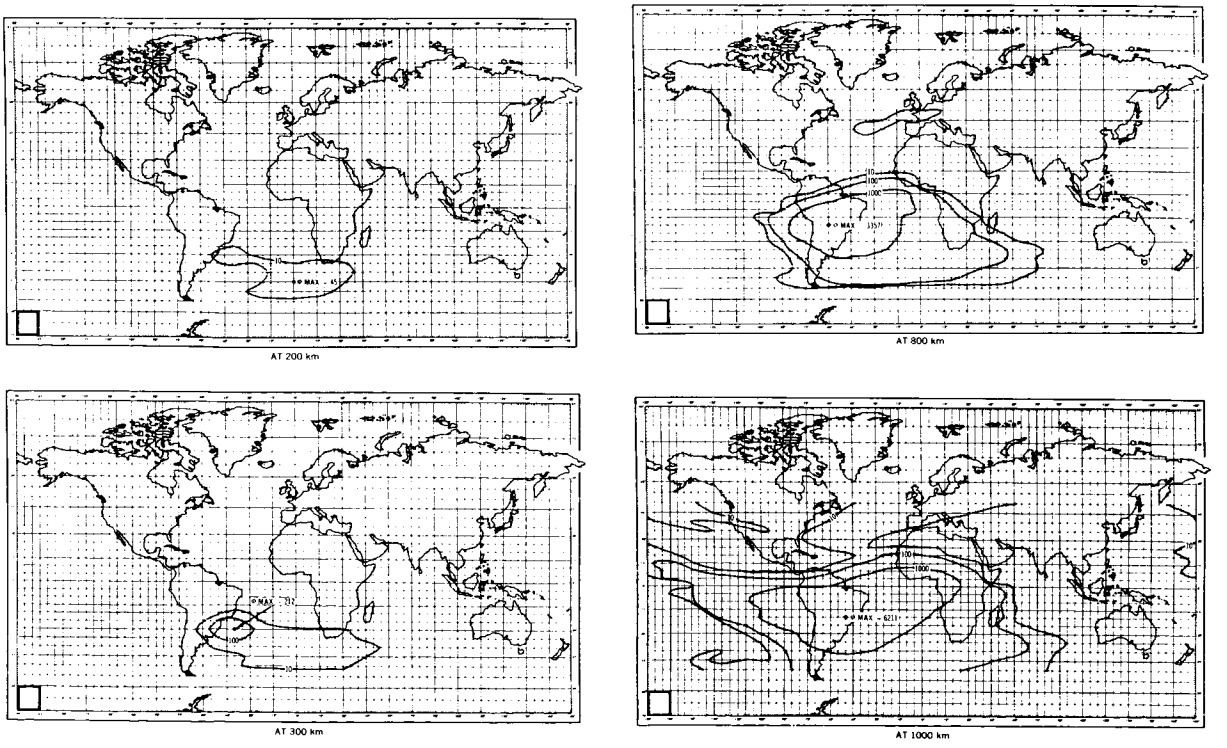


FIGURE 7.—Electron flux maps at different altitudes above the earth's surface. Flux is in units of 10^5 electrons/cm^2-sec.

TABLE 1.—*Calculations on Fluxes Encountered by Satellites Moving Through the Artificial Radiation Belt*

Datum	Ariel I	Traac and Transit IV–B	Telstar I	Tiros V	Orbiting solar observatory I (1962 ζ1)	Relay*	Polar orbiting geophysical observatory (Pogo)*
Perigee (km)	390	960	952	590	552	1343	257
Apogee (km)	1210	1106	5660	971	594	5555	931
Inclination (degrees)	54	32	45	58	33	50	90
Altitude (km) at 30°S latitude, 30°W longitude.	1067	1000	5138 1758	963	594	4371 1516	804
Calculated r/day outside vehicle	110,000	180,000	800,000	46,000	27,000	1.1×10^{6}	22,000
Code output from machine	$\frac{3.1\times10^{9}}{4}$	$\frac{5.0\times10^{9}}{4}$	$\frac{2.2\times10^{10}}{4}$	$\frac{1.28\times10^{9}}{4}$	$\frac{7.5\times10^{8}}{4}$	$\frac{3.0\times10^{8}}{4}$	$\frac{6.2\times10^{8}}{4}$
Length of machine run in satellite days	4	4	4	4	4	4	4
Electrons/cm²-day	2.8×10^{12}	4.5×10^{12}	2.0×10^{13}	1.15×10^{12}	6.8×10^{11}	2.7×10^{13}	5.6×10^{11}
Protons/cm²-day	—	—	—	—	—	—	—

*To be launched.

TABLE 1.—(Continued)

Datum	1000 km polar orbit†	800 km polar orbit	Orbiting astronomical observatory*	SERB*	MA-7 (1962 τ1)
Perigee (km)	1000	800	802	278	160
Apogee (km)	1000	800	817	16,668	264
Inclination (degrees)	90	90	31	17.0	33
Altitude (km) at 30°S latitude, 30°W longitude	1000	755	810	—	261
Calculated r/day outside vehicle	80,000	27,000	—	—	80‡
Code output from machine	$\frac{2.2\times10^9}{4}$	$\frac{1.04\times10^9}{4}$	$\frac{2.2\times10^9}{4}$	$\frac{1.4\times10^{10}}{4}$	3.45×10^6
Length of machine run in satellite days	4	4	4	4	0.4 (9.5 hr)
Electrons/cm²-day	2×10^{12}	9×10^{11}	2×10^{12}	1.2×10^{13}	0.24×10^{10}‡
Protons/cm²-day	—	—	—	—	—

†Similar to Nimbus, the Fixed Frequency Topside Sounder, the Swept Frequency Topside Sounder, and the Polar Ionosphere Beacon (which have been or will be launched).

‡Six orbits only.

TABLE 2.—*Orbital Elements of Various Space Vehicles*

Element	Ariel I	Traac and transit IV–B	Telstar I	Tiros V	OSO I	Relay
Epoch (days, hours, min, sec)	190,9,0,0	190,4,3,46.506	191,8,51,0	190,9,0,0	190,9,0,0	305,0,0,0
Semimajor axis (earth radii)	1.1254	1.1618	1.5182	1.1224	1.0900	1.5407
Eccentricity	0.05714	0.009922	0.2430	0.02663	0.003012	0.2143
Inclination (degrees)	53.866	32.423	44.803	58.102	32.855	50.0003
Right ascension of ascending node (degrees)	−24.881	96.434	−156.222	−75.536	154.502	163.708
Argument of perigee (degrees)	−9.2537	−51.6890	164.811	118.014	139.136	−167.526
Mean anomaly (degrees)	−86.8833	0.0001	1.1684	−194.11968	−164.5453	7.8219

Element	Pogo	1000 km polar orbit	MA–7	800 km polar orbit	OAO	SERB
Epoch (days, hours, min, sec)	82,3,55,32.101	190,9,0,0	268,14,0,0	190,9,0,0	153,0,0,0	303,10,0,0
Semimajor Axis (earth radii)	1.0931	1.1568	1.0331	1.1254	1.1270	2.3284
Eccentricity	0.04830	0.1490×10^{-7}	0.008552	4.4703×10^{-7}	0.001074	0.5518
Inclination (degrees)	90.001	90.000	32.546	90	30.982	17.0
Right ascension of ascending node (degrees)	−73.806	−158.175	75.069	−158.175	38.6643	17.2908
Argument of perigee (degrees)	−19.408	−180.000	78.188	−180.000	−68.8243	134.6735
Mean anomaly (degrees)	2.1956	0.0000	7.6908	0	36.7572	0

produce the observed power supply damage (private communication from A. Franta, Goddard Space Flight Center). This would be caused by about 10^{13} electrons/cm^2 according to figure 8. About seven days after the nuclear explosion, this flux would have been achieved (Table 1 gives 2.8×10^{12} electrons/cm^2-day for Ariel I, of which half hit the face of the cells). The Ariel I power supply started malfunctioning in $3\frac{1}{2}$ days. This is quite good agreement.

Traac and Transit IV–B also had blue sensitive p-on-n solar cells, but it would take 3×10^{14} electrons/cm^2 to cause malfunction, because the cells were lower efficiency cells (private communication from R. Fischell, Applied Physics Laboratory). Table 1 gives 4.5×10^{12} electrons/cm^2 encountered per day. Half of these electrons hit the face of the cells. Traac stopped transmitting in 36 days and Transit IV–B in 24 days. Using 30 days as the average, we get a total encountered flux of 0.7×10^{14} electrons/cm^2, only in fair agreement with that required to produce damage.

Telstar I used the much more damage resistant

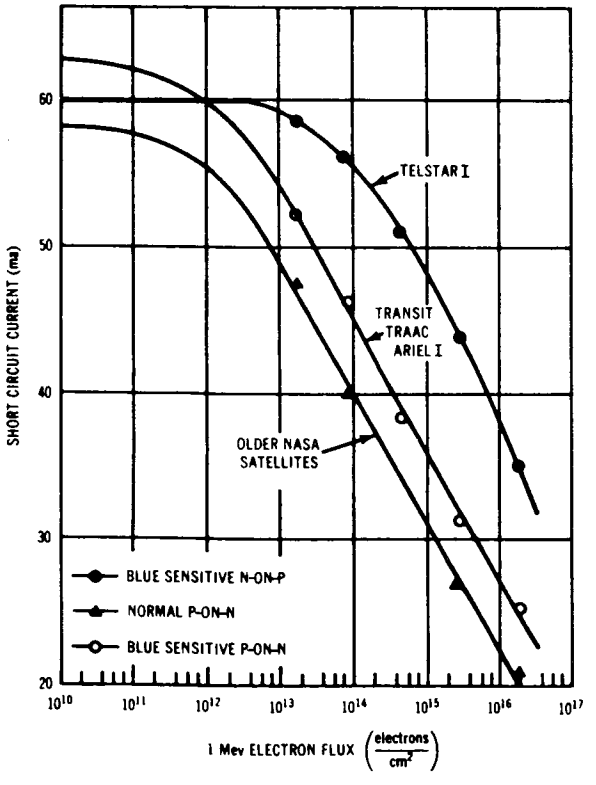


FIGURE 8.—Solar cell damage curves.

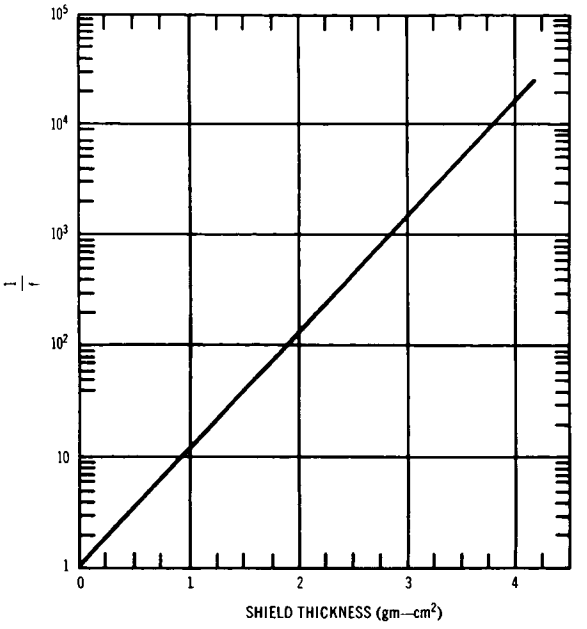


FIGURE 9.—The fraction of fission electrons that penetrate different shield thicknesses.

n-on-p cells, because it was to routinely fly through the inner radiation belt protons. Even with the artificial radiation belt, its power supply lifetime is expected to be considerably longer than 1 year.

The Telstar I solar cells are degrading at a rate that would be produced by 6×10^{12} electrons/cm^2-day of 1 Mev hitting the bare cells (private communication from W. Brown, Bell Telephone Laboratories). This corresponds to about 1.8×10^{13} electrons/cm^2-day incident on the outside of the 30 mil sapphire covers. Our calculations give $\frac{1}{2}\times2\times10^{13}=1\times10^{13}$ electrons/cm^2-day hitting the cells. The observed solar cell degradation on Telstar I should be somewhat more than that calculated from the artificial electron belt, because slow proton damage probably contributes somewhat to the degradation (private communication from W. Brown, Bell Telephone Laboratories).

We have neglected the enhanced early time effects here on all the exposed satellites. An appreciable part of the encountered flux may have been encountered in the first few days. For the first week after the explosion the flux was higher than that given in figures 5 and 6.

Injun, Tiros V, and other satellites continue to

TABLE 3.—*Flux per Orbit for a Mercury Capsule at an Altitude of* **264** *km*

Orbit	Flux (electrons/cm²)
1	5.0×10^{6}
2	2.1×10^{7}
3	4.8×10^{7}
4	2.9×10^{8}
5	6.4×10^{8}
6	1.4×10^{9}

function. Injun has a low duty cycle and Tiros V shows some solar cell degradation. Film badge dosimeter measurements have been made on several space flights. About 10 r/day was measured inside 1.5 gm/cm² of shielding. In order to compare this radiation dose with the predictions in Table 1, correction must be made for shielding. To do this we perform a calculation like that done for the Traac GM counter to get f, the fraction of electrons that penetrate the wall. Values of f have been calculated for different thicknesses of shield by using the relationship $R=0.526\,E-0.094$ and the associated rough-straggling transmission curves in figure 3. Figure 9 shows a plot of $1/f$ as a function of shield thickness. This is really only true for Al but for lack of better information we will use it for other materials too. For 1.5 gm/cm² we get $f=1/50$ for normal incidence particles. To correct for a distribution of incidence angles we will say roughly that about half as many get through. Also, 2π steradians are covered by a much thicker shield so that the total factor $f'=1/200$. This would mean that 10 r/day$\times$200=2000 r/day were incident on the outside of the vehicle. This agrees to within a factor of 2 with the calculated vehicle-encountered flux.

MANNED FLIGHT

For a Mercury capsule orbit with an apogee of 264 km the total flux encountered in six orbits would be 0.24×10^{10} electrons/cm² outside the vehicle (Table 3). If the apogee is lowered by 30 km (to 234 km) the total flux for 6 orbits is reduced to 0.17×10^{10} electrons/cm². If the apogee is raised by 30 km (to 294 km) the total flux for 6 orbits is increased to 0.45×10^{10} electrons/cm².

PARTICLE TIME HISTORIES

One of the important problems to answer about the new belt is how long it will last. The currently intense regions will last a number of years, according to present indications. At low altitudes the fluxes have already decayed a lot. According to Ariel I and Traac data, outside the 10^5 contour of the B–L plot in figure 6 the fluxes decayed several orders of magnitude in a few days.

Injun has noted some decay at 1000 km (private communication from B. O'Brien, State University of Iowa). At $L=1.18$ and $B=0.191$ there is a decay factor of about 2 from +10 to +1000 hr. For the same L and time interval for $B=0.206$ there is a decay factor of 4 and for $B=0.214$ there is a decay factor of 60. Injun saw no marked change in flux as a result of a modest size magnetic storm.

The only decay process we understand well enough to calculate is coulomb scattering. Particle time histories have been calculated for coulomb scattering and characteristic times determined (Reference 5). The time to reach a scattering equilibrium (which is also about the time for this equilibrium to decay to $1/e$ intensity) for different L values is listed in Table 4. Welch, Kauffman, et al.[5] first calculated these for solar maximum atmospheric densities and now, assuming that the density is less by a factor of 10, we get the values in Table 4.

The air densities are not well known and the calculated times may be wrong by a factor of 5 or more. The calculated variation with L, however, should be fairly good. The Injun data show that the calculated times are of the right order of

TABLE 4.—*Time Until Scattering Equilibrium for Different Values of L*

L	Calculated τ (days)	Measured τ
1.20	10	~1 month
1.25	150	—
1.30	1500	—
1.35	~3000	—
1.40	~10,000	—

magnitude. The times show that the high flux region should last even through the next solar maximum if coulomb scattering is the principle loss process.

REFERENCES

1. O'Brien, B. J., Laughlin, C. D., and Van Allen, J. A., "Geomagnetically Trapped Radiation Produced by a High-Altitude Nuclear Explosion on July 9, 1962," *Nature*, **195**(4845): 939–943, September 8, 1962.
2. Marshall, J. S., and Ward, A. G., "Absorption Curves and Ranges for Homogeneous β-Rays," *Can. J. Res.*, **15A**, 39–41, March 1937.
3. McIlwain, C., "Coordinates for Mapping the Distribution of Magnetically Trapped Particles," *J. Geophys. Res.*, **66**(11): 3681–3691, November 1961.
4. Rosenzweig, W., Gummel, H. K., and Smits, F. M., "Solar Cell Degradation Under 1 Mev Electron Bombardment," submitted to *J. Appl. Phys.*
5. Welch, J. A., and Kauffman, R., et al., "Scattering Loss of Fission Beta Particles from High Altitude Explosions," Air Force Special Weapons Center Report 2–0038, August 1962.

Appendix A

SHIELDING AND RADIATION DOSES

Included here for the sake of completeness are some crude calculations on shieldings and dosages.

One consideration that is important in some shielding calculations is bremsstrahlung. The doses delivered by the x-rays made by bremsstrahlung will be larger than the direct electron doses for large shield thicknesses.

The fraction of the energy of an electron that goes into bremsstrahlung may be calculated from:*

$$E_{\text{brem}}/E_{\text{ion}} = ZE^2/1600.$$

where Z is the atomic number of the material involved. For the fission energy spectrum the average energy is about 1 Mev:

$\frac{E_{\text{brem}}}{E_{\text{ion}}} =$	C	Al	Fe	Pb
	0.004	0.008	0.015	0.050

The energy spectrum of the x rays will be something like that in figure A1. There will be a very few x rays up to 8 Mev, but not many over 2 or 3 Mev. The low energy x rays (below about 100 kev) will be absorbed in the shielding. This will remove about half the total energy in the x rays. The resultant transmitted energy spectra will have a peak at about $\frac{1}{2}$ Mev (figure A1). The x rays transmitted through the shield will be quite penetrating. Their mean free path will be roughly 20 gms/cm^2. This means two things. First, they will be hard to absorb, and therefore it will take a lot more shielding to absorb them. Second, because they are hard to absorb, they will not be counted efficiently by a particle counter and also will result in less radiation dose.

We can now calculate crudely the counting efficiency of the Injun (1961 α2) 213 GM counter. From figure 9 of the body of this report we see that it would only count about 1/20,000 of the fission electrons directly. But we find that 0.05 of the energy is converted to bremsstrahlung, of which half is absorbed in the shield. The mean energy of these x rays will be about $\frac{1}{2}$ Mev. A normal GM counter will detect these x rays with about 1 percent efficiency. This gives

$$(0.05)(0.01) = 1/2000.$$

for the fraction of the electrons counted via bremsstrahlung. This calculation is not very accurate but it does show that the Injun counter counts electrons via bremsstrahlung with about the observed efficiency.

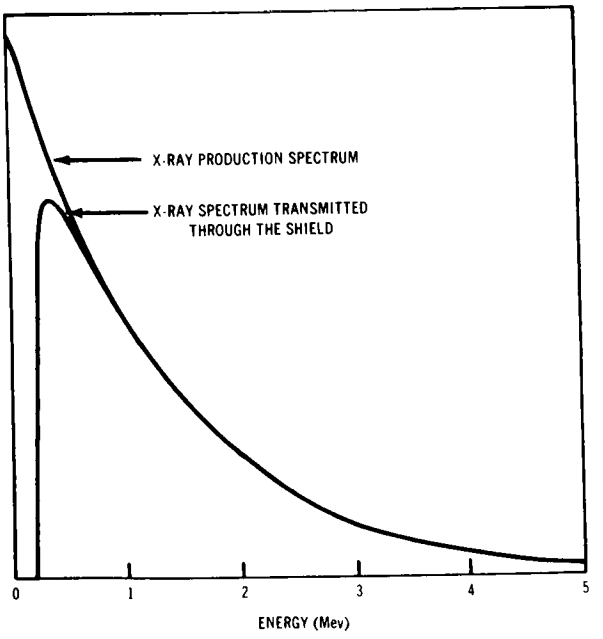


FIGURE A1—A crude bremsstrahlung X-ray energy spectrum.

*Fermi, E., "Nuclear Physics," A course given by Enrico Fermi at the University of Chicago, Notes compiled by Orear, J., Rosenfeld, A. H., and Schluter, R. A., University of Chicago Press, revised edition, 1950.

TABLE A1.—*Radiation Dose per Orbit for a 6 Orbit Flight on the MA–7 Orbit*

Orbit	Dose (*r*)
1	0.003
2	0.01
3	0.03
4	0.1
5	0.3
6	0.6

MANNED FLIGHT

The effects of the new radiation belt on manned flights must be considered. For the Mercury project the total flux that would be encountered for a six orbit mission with the MA–7 (1962 $\tau 1$) orbit is 0.24×10^{10} electrons/cm^2 outside the vehicle, or 80 r (3×10^7 electrons/cm^2 = 1 r). The shielding of the vehicle is such that about 1 percent of this dose is delivered to the astronaut, about 1 r.

The mercury dose is almost all received in the South Atlantic "hot spot" (see figure 8 of the body of the report) and occurs mainly on orbits 4, 5, and 6. The breakdown of the 1 r dose inside the capsule by orbits is given in Table A1.

ENERGETIC PARTICLES IN THE INNER VAN ALLEN BELT

WILMOT N. HESS

Goddard Space Flight Center

A survey is given of the high-energy particle populations in the inner radiation belt. The experiments which have given information about particles are discussed and the best experimental information about particle fluxes and spectra is presented. Calculations are given which indicate what particle sources and loss processes are most important. The role of neutrons in making the inner belt is considered in detail and the need for particle acceleration and/or other sources of particles is shown.

INTRODUCTION

The satellite Explorer I (1958α) launched with a Geiger–Müller counter on board discovered a region of high count rate starting at an altitude of about 1000 km. This was unexpected—in fact, it was suggested that the counter might have malfunctioned. But results from Explorer III (1958 γ) demonstrated that the effect was real. Van Allen, who had conducted both experiments,[1] realized very soon that the measured high count rates were due to charged particles trapped in the earth's magnetic field.[2] Störmer had worked extensively on this general subject[3] and even calculated orbits of trapped particles[4] years earlier, but the actual existence of trapped particles had not been suggested in this work. Ideas about the existence of a terrestrial ring current had also essentially included the idea of trapped particles.[5]

At the same time that these experiments in space were going on, experiments with trapped particles were going on in various laboratories. Project Sherwood is an attempt by the Atomic Energy Commission to make a controlled thermonuclear reaction on a small scale[6] by confining charged particles in a magnetic field. Christofilos, who was working on Project Sherwood, extrapolated the laboratory idea to earth scale and suggested the possibility of trapping a large number of charged particles in the earth's magnetic field by using a nuclear explosion to inject the particles. This idea was carried out in the Argus Experiment[7] and demonstrated experimentally that charged particles could really be trapped in the earth's field. It had even been suggested in planning the Argus Experiment that a natural radiation belt, populated by the decay of neutrons escaping from the earth's atmosphere might exist; but this idea did not get wide distribution.

Results obtained from Sputnik III (1958 δ2) in May 1958 confirmed the existence of the trapped radiation with measurements up to 1800 km.[8,9]

Explorer IV (1958 ϵ) was launched in July 1958 with instrumentation to study the natural radiation belt[10] and to study the artificial belt produced by Argus.[11] A map of the radiation intensities up to an altitude of 2200 km and a range spectrum of the charged particles were obtained by Explorer IV.

Pioneer III (1958 θ1) in December 1958[12] and Pioneer IV (1959 ν) in March 1959[13] made isolated passes more or less radially outwards through the inner belt. More recently, Injun (1961 o2) and Explorer VII (1959 ι1) and several Discoverer satellites have given information about the inner belt. These, with various rocket shots, complete the list of experiments performed.

In this report, an attempt is made to survey the experiments that have given information

TABLE 1.—*Characteristics of Particles at 2000 km Altitude Near the Equator*

Particle	R_c (cm)	τ_c (sec)	τ_B (sec)	τ_R (min)
50 kev electron	5×10^{3}	2.5×10^{-6}	0.25	690
1 Mev electron	3.2×10^{4}	7×10^{-6}	0.10	53
1 Mev proton	1×10^{6}	4×10^{-3}	2.2	32
10 Mev proton	3×10^{6}	42×10^{-3}	0.65	3.2
500 Mev proton	2.5×10^{7}	6×10^{-3}	0.11	0.084

about the inner belt and also the calculations that have been made dealing with energetic particles in this area. One subject that will receive special attention is neutrons in space and neutron decay, because it is well established that neutrons play a major role in the production of protons and also probably of electrons in the inner belt.

MOTION OF PARTICLES IN A DIPOLE FIELD

The general problem of charged particle motion in a dipole field is complicated.[3] Fortunately, for radiation belt particles, an approximation can be used which simplifies the situation considerably. Alfvén[14] introduced the idea of the *guiding center* of a particle. The particle motion in this case is described in terms of: (1) a rapid gyration about a *guiding center* with a cyclotron period τ_c and radius of gyration R_c, and (2) motion of this guiding center along a magnetic line of force. The motion along the line is periodic too. The particle is reflected by the converging magnetic field near the earth and bounces back and forth in the exosphere with a bounce period τ_B. There is another motion,[15] a slow drift in longitude around the earth with a period of revolution of τ_R. Particles at 2000 km altitude near the equator will have the characteristics listed in Table 1. Because the three periods are so different, the particle motion is separable into these three components. If the cyclotron radius of the particle R_c becomes comparable to the diameter of the earth, the motion is not separable; however, even for 1 Bev protons this condition does not occur.

We can understand the particle's bouncing motion in the following way.[16] A static magnetic field does not work on a particle; therefore, the flux linking the orbit of a particle rotating about a field line is constant. If $d\Phi/dt=0$, the particle's energy would change. Thus,

$$\Phi = B\pi R_c^2 = \text{constant}, \tag{1}$$

and we can write for the particle's perpendicular energy

$$E_\perp = \frac{mV_\perp^2}{2} = \frac{m\omega^2 R_c^2}{2} = \frac{mR^2}{2}\left(\frac{eB}{mc}\right)^2 = \frac{e^2B^2R_c^2}{2mc^2}. \tag{2}$$

Substituting this into Equation 1, we get

$$\Phi = \text{constant} = 2\pi mc^2 E_\perp/e^2B; \tag{3}$$

therefore,

$$E_\perp/B = \text{constant} = \mu. \tag{4}$$

The constant μ is the magnetic moment of the particle's motion around the field line.

From Equation 4 we see that

$$(\sin^2\alpha)/B = \text{constant}. \tag{5}$$

The particle will move into a region of increasing B until $\sin\alpha=1$ when it must turn around. It then moves out of the high field region and repeats the process at the other end of the field line for the same value of B. The angle α is the particle's pitch angle.

The drift in longitude of a charged particle results from a force on the particle perpendicular to the field lines and lying in the plane through the center of the earth containing the earth's

axis. The magnitude and direction of this drift velocity can be obtained from the cyclotron equation

$$\mathbf{R} = \frac{mc}{eB^2}(\mathbf{v}\times\mathbf{B}) = \frac{c}{eB^2}(\mathbf{p}\times\mathbf{B}) . \quad (6)$$

If a force $f_\perp$ acts perpendicular to B for a time Δt, there is a change of momentum of

$$\Delta p = f_\perp \Delta t; \quad (7)$$

this results in a displacement of the guiding center of the particle of

$$\Delta\mathbf{R} = \frac{c}{eB^2}(\Delta\mathbf{p}\times\mathbf{B}). \quad (8)$$

Differentiating equation 8 with respect to time gives

$$\mathbf{v}_D = \frac{d\mathbf{R}}{dt} = \frac{c}{eB^2}\left(\mathbf{f}_\perp\times\mathbf{B}\right) . \quad (9)$$

This drift velocity is perpendicular to both $f_\perp$ and B; therefore, if $f_\perp$ lies in the plane through the center of the earth containing the earth's axis, the drift velocity will be azimuthal—a drift in longitude.

One force that produces a drift is due to the gradient of the earth's magnetic field:

$$\mathbf{f}_\perp = \mu\nabla\mathbf{B} = \frac{3\mu B}{R} . \quad (10)$$

A second force that produces a drift is the centrifugal force on a particle due to the curvature of field lines:

$$\mathbf{f}_\perp = \frac{mv_{11}^2}{R_l} . \quad (11)$$

Combining these gives a drift velocity

$$v_D = \frac{1}{\omega_c R_c}\left(v_\parallel^2 + \frac{1}{2}v_\perp^2\right) . \quad (12)$$

where ω_c and R_c are the cyclotron frequency and radius, respectively.

There is also a possibility of radial drift of particles due to line exchange[17] or other effects.[18] but these drifts do not seem to be important for high-energy trapped particles as was demonstrated by the spatial stability of the Argus electrons.

ADIABATIC INVARIANTS

Associated with the motion of particles in a dipole field are three constants. They are actually only adiabatic constants—in other words they are constant unless magnetic fields change rapidly. We have shown the magnetic moment μ to be a constant of the motion, but it is not constant if fields change in times short compared to τ_c or in distances short compared to R_c.

The second adiabatic[19] invariant I is called the integral invariant,

$$I = \frac{1}{v}\int_{m_1}^{m_2} v_{11}\, dl , \quad (13)$$

where the integral is taken along a field line between the two mirror points. This quantity is related to the Hamilton action integral and is violated if changes occur in times short compared to τ_B.

The third invariant is the flux invariant.[20] When this invariant holds, the magnetic flux linked by the particle's orbit is constant. This is violated if field changes occur in times short compared to τ_R.

We can understand the motion of a particle in the earth's field rather well by considering only the constancy of μ and I, From μ we know that the particle mirrors at a particular value of B. From I we can determine which field line the particle will travel as it drifts around the earth. The combination of these two defines a surface around the earth, resembling the surface of a pitted olive, on which the particle will travel.

SOURCES OF THE BELT PARTICLES

Cosmic Rays

A flux of about 2 particles/cm^2-sec of galactic cosmic rays reaches the earth regularly. One possibility about the radiation belts is that they represent quasi-trapped particles,[21] that is, they consist of particles on certain special Störmer orbits that can stay near the earth for a long time and then finally move out of the earth's field and escape. Estimates of the intensity of the radiation belt that would be produced this way show this to be a small effect. A flux increase by a factor of 10^4 or more over the galactic cosmic rays would be needed to produce the radiation

belt fluxes. The increase obtained by this quasi-trapping is nowhere near 10^4.

Another possibility is that cosmic rays might produce trapped particles by interacting with the very thin atmosphere at very high altitude. We can estimate the importance of this source. If we take an atmospheric density of 10^5 atoms/cm^3 of oxygen and a cross section for producing high-energy charged particles of 0.2×10^{-24} cm^2, we get a source strength S of

$$S=\left(\frac{2\text{ protons}}{\text{cm}^2\text{-sec}}\right)\left(\frac{10^5\text{ atoms}}{\text{cm}^3}\right)\left(\frac{0.2\times10^{-24}\text{ cm}^2}{\text{atom}}\right)$$

$$=\frac{4\times10^{-20}\text{ protons}}{\text{cm}^3\text{-sec}}.$$

This also is negligible when compared to other source strengths.

Still another possible source of the belt particles is splash albedo protons produced by interactions of the high-energy cosmic rays with the atmosphere. We know that the particles must return to roughly their birth altitude or even lower in order to mirror at the other end of their line of force. They must therefore encounter a rather thick atmosphere and be rapidly lost, so this source contributes little, if anything.

Mu Mesons

High-energy collisions involving cosmic rays produce π mesons. These decay in about 10^{-8} sec to μ mesons which in turn decay in 2×10^{-6} sec to electrons having energies up to 50 Mev. If the decay occurs at a high enough altitude, the electron can be trapped. A μ meson, to travel 1000 km, must live 0.003 sec. Even if it lives 5 half-lives or 10^{-5} sec, it must have a relativistic time dilation factor of $\gamma=0.003/10^{-5}=300$ in order to travel this distance. Thus it must have an energy on the order of 300×106 Mev$=30$ Bev. There are very few 30 Bev μ mesons made traveling upwards out of the atmosphere, and this source can also be neglected.

Solar Wind

There are about 10 protons/cm^3 in the kev energy range striking the magnetosphere;[22] this is a large flux of particles. If they could be brought into the magnetic field of the earth, they could be an important source. There are several possibilities here. Particles may be injected at the field nulls near the poles.[23] Field convection processes may bring them into the field;[24] Taylor instabilities may also help them get in. All of these are possible but none of them are understood well enough to be evaluated. It would appear that some processes of this sort may produce low energy particles in the outer belt efficiently. For want of any specific information about these processes we will not consider them further.

Neutrons

Neutrons are made by cosmic ray protons colliding with oxygen and nitrogen nuclei in the atmosphere. Several neutrons are made in the nuclear cascade in the atmosphere. A 5 Bev cosmic ray proton will produce about 7 neutrons in the atmosphere. About 25 percent of all neutrons produced this way diffuse out into space; this leakage flux of neutrons out of the atmosphere is about 0.5 neutrons/cm^2-sec at the equator and about 5 neutrons/cm^2-sec at the pole. The neutron flux in space close to the earth at low latitudes has been measured[25] to be about 1.0 neutrons/cm^2-sec; this is in reasonable agreement with the calculated fluxes. Few neutrons reach the earth from the sun or more distant space because the free neutron is radioactive with a mean life of only 1000 sec. It decays by the reaction

$$n\rightarrow p+e+\bar{\nu}.$$

The antineutrino $\bar{\nu}$ does not interest us here, but the proton and electron are both important in forming the radiation belt.

We need to know how many neutrons of what energies decay at different places in space to produce protons and electrons. First, we must know the neutron energy spectrum $\Phi_n(E,\ R,\ \lambda)$ in space at all radii R and latitudes λ. This has been calculated[26] from a knowledge of the neutron energy spectrum inside the atmosphere[27] and is shown in figure 1. The flux decreases as a function of height above the earth. This is especially true for low-energy neutrons. Neutrons of less than $\frac{2}{3}$ ev are trapped by the earth's gravitational field so that they essentially all decay in space

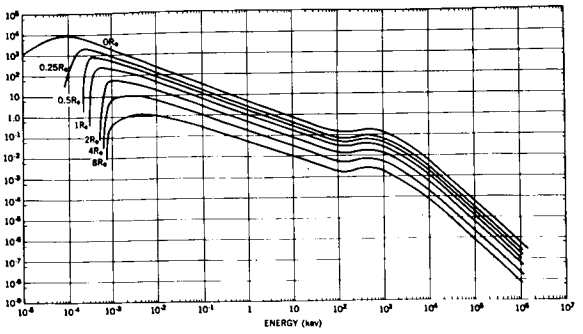


FIGURE 1.—Albedo neutron energy spectra in space at different distances above the geomagnetic equator on December 5, 1959. The zero R_e curve is for the top of the atmosphere, roughly 100 km. Reproduced with permission from references 26 and 27.

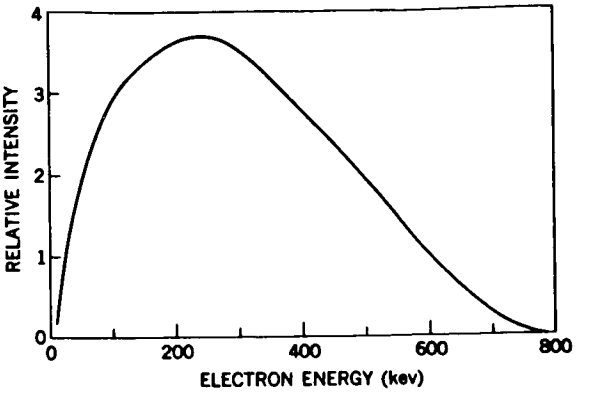


FIGURE 2.—The electron-energy spectrum from neutron β decay. To obtain electron flux multiply the intensity by the electron velocity.

near the earth, while only about 1 percent of neutrons of 1 Mev energy decay near the earth. The density of neutrons decaying is given by

$$\frac{dn}{dV}(E,R,\lambda) = \frac{1}{\gamma v(E)\tau_n}\Phi_n(E,R,\lambda) , \quad (14)$$

where v is the neutron velocity and τ_n the neutron mean life.

The electrons made by the neutron decay will have the β decay spectrum shown in figure 2. The electron's energy will be essentially unchanged by the neutron's kinetic energy. To demonstrate this, consider a 10 Mev neutron, which has a velocity of about 4×10^9 cm/sec. When an electron of about 30 kev energy is produced by the neutron decay it has a velocity of about 2×10^{10} cm/sec. When the neutron's velocity is compounded with this velocity relativistically it changes by, at most, 10 percent. Almost all neutrons have energies less than 10 Mev, so the effect of the neutron's motion in changing the electron's energy can be neglected. Therefore, to determine the total electron source strength S_e for the radiation belt due to neutron decay, we simply add up all neutron decay events:

$$S_e(R,\lambda) = \int \frac{dn}{dV}(E,R,\lambda)\,dE = \int \frac{1}{\gamma v \tau_n}\Phi(E,R,\lambda)\,dE . \quad (15)$$

This electron source strength S_e is shown in figure 3. We see there are from 10^{-13} to 10^{-12} electrons/cm^3-sec made by this source in space near the earth.

The situation is different when we consider the protons made by neutron decay. The kinetic energy of the proton is very nearly the kinetic energy of its parent neutron. The electron kinetic energy is supplied by the mass difference of the neutron and proton. The recoil energy given the proton by the electron is only about 100 ev; so, if we consider protons above about 10 kev, we can accurately take the proton's energy and direction of motion to be that of the parent neutron. Because of this the decay density energy spectrum shown in figure 4 is also the proton source energy spectrum $S_p(E, R, \lambda)$ from about 10 kev up,

$$S_p(E,R,\lambda) = \frac{dn}{dV}(E,R,\lambda) . \quad (16)$$

There are about 6×10^{-15} protons/cm^3-sec of $E>10$ Mev produced in space near the earth. The proton source near the earth at the equator has been evaluated from figure 1 to be

$$S_p(E,R_e,0) = \frac{0.8E^{-2.0}}{\gamma v \tau_n} . \quad (17)$$

We have a quantitative picture of the neutron-

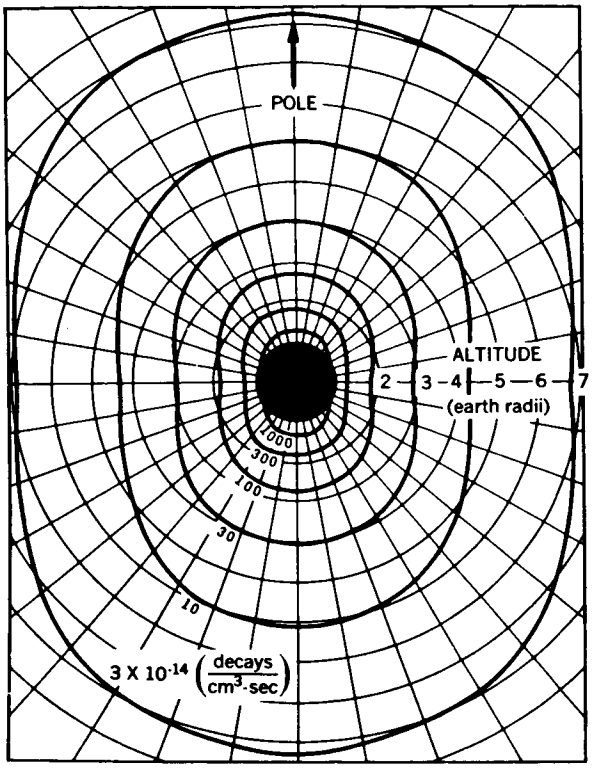


FIGURE 3.—Values of the neutron decay density dn/dV summed over neutron energy at different attitudes and magnetic latitudes. Reproduced with permission from reference 26.

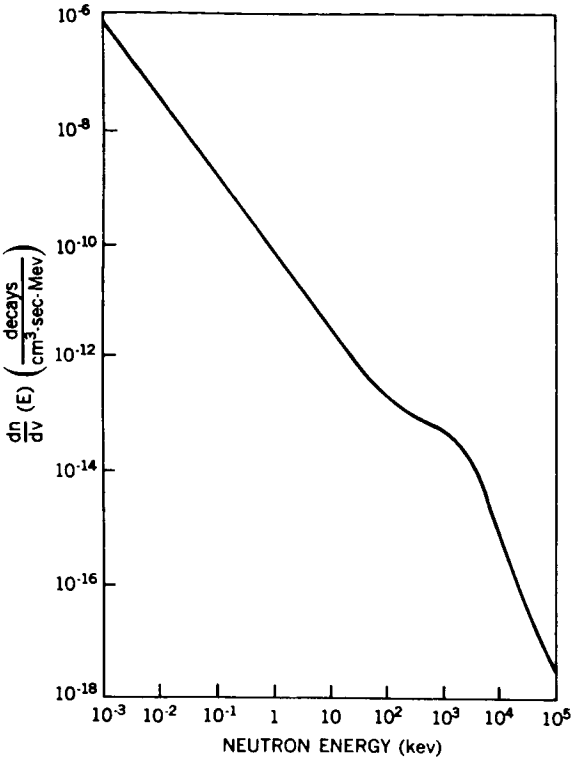


FIGURE 4.—The neutron decay energy spectrum $dn/dV(E)$ at 1500 km on the Equator.

decay proton and electron sources. These are an important source of the radiation belts.

LOSS PROCESSES

Protons

There are at least three ways in which protons are lost from the radiation belt. In the inner belt, high-energy protons are removed most rapidly by slowing down until they reach about 100 kev. The amounts of oxygen R that must be traversed to stop protons of various energies are listed in Table 2.

TABLE 2.—*Ranges of Protons in Oxygen*

Energy (Mev)	Range (gm/cm²)
0.1	6×10^{-5}
1	0.003
10	0.14
100	8.6

Below 100 kev, the protons are more rapidly lost by charge exchange[28] with slow protons by the reaction

$$\underline{p} + H \rightarrow p + \underline{H},$$

where the bar refers to the fast particle. At 50 kev, the cross section for charge exchange[29] is $\sigma_{cH} = 2 \times 10^{-16}$ cm², and the charge exchange lifetime τ_{cH} is given by

$$\tau_{cH} = \frac{1}{\sigma_{cH}\, n(H)\, \nu} = \frac{1}{\left(2 \times 10^{-16}\ \text{cm}^2\right)\left(10^4\ \frac{\text{atoms}}{\text{cm}^3}\right)\left(3 \times 10^8\ \frac{\text{cm}}{\text{sec}}\right)} = 1600\ \text{sec} \qquad (18)$$

where $n(H)$ is the atomic density of hydrogen, taken here to be 10^4 atoms/cm^3. This time is less than the slowing down time above approximately 1000 km altitude, so the charge exchange process will dominate here.

In the outer belt, protons are lost rapidly by some additional process. Probably this involves the breakdown of the magnetic moment invariant. Various possibilities along this line are considered later.

Electrons

Different processes are responsible for the loss of electrons than for protons. Because electrons are lighter, they scatter more easily and are lost in the inner belt principally by coulomb scattering into the loss cone rather than by slowing down. This can be seen by comparing R, the range of the particle, and D, the amount of material necessary to scatter out. The range of a 1 Mev electron, from Feather's rule, is

$$R = 0.54E - 0.13 = 0.41 \frac{\text{gm}}{\text{cm}^2} \cdot \quad (19)$$

The value of D can be obtained by using the scattering formula[30]

$$\theta^2 = \frac{7000\,D}{E^2} \cdot \quad (20)$$

where θ^2 is the mean angle of scattering, D is the path length of air (at S.T.P.) traversed in cm, and E is the particle energy in kev. Taking $\theta = 0.5$ rad as the necessary scattering angle to lose the particle, we find

$$D = 0.04 \frac{\text{gm}}{\text{cm}^2} \cdot$$

This shows the electron will be scattered out of the inner belt before it slows down.

HIGH-ENERGY PROTONS

Experiments on the Energy Spectrum

The first experiment performed in the radiation belt that unambiguously identified the partices which were counted involved flying a stack of nuclear emulsions on the Atlas rocket.[31] The emulsion stack was recovered and developed and the nuclear tracks were read. The range and ionization of the particles were measured, the particles identified and their energies determined.

Protons of $E > 75$ Mev and electrons of $E > 12$ Mev could get through the 6 gm/cm^2 shielding into the nuclear emulsions. No electrons were found, but a large number of protons were found. The energy spectrum of protons measured by Freden and White on a later flight[32] is shown in figure 5. Other experiments[33,34,35,36] have shown very similar energy spectra and intensities of protons and have extended the data down to lower energies. Below about 40 Mev the energy spectrum becomes quite flat and shows a slight dip at about 20 Mev. The inner-belt proton flux seems quite constant in time, varying by less

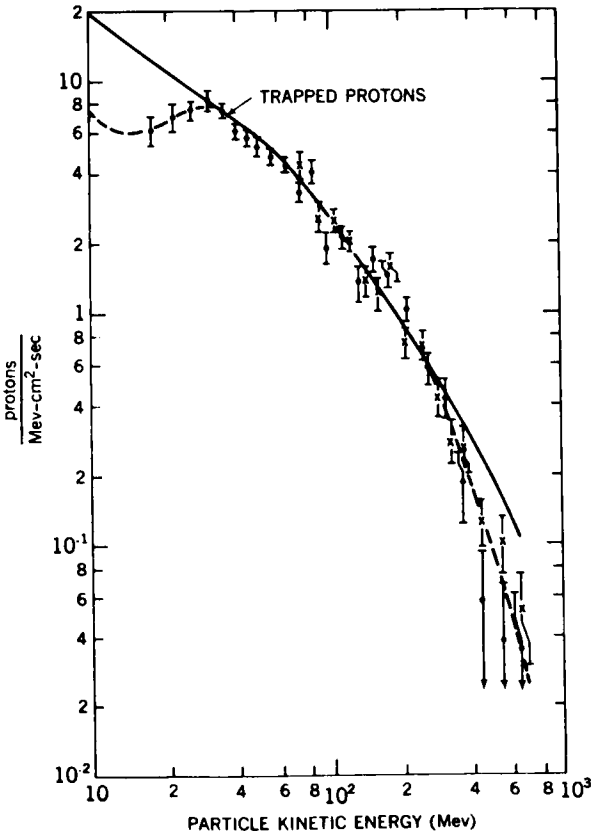


FIGURE 5.—The spectrum of trapped high-energy protons at 1100 km measured by Freden and White. Reproduced with permission from reference 32. The solid dots are data from a flight on October 13, 1960, and the crosses are from a flight in April 1959, normalized to the other at 100 Mev. The solid curve is the theoretical proton curve. The dashed curve is a fit to the high-energy data.

than a factor of 2 for several flights. This indicates that the particle lifetime is quite long.

Calculation of the Energy Spectrum

We can get a quantitative picture of the flux and energy spectrum of the inner-radiation-belt protons produced by neutron β decay by considering the conservation of particles. The continuity equation in energy space can be written[37]

$$\frac{dN(E)}{dt} = S_p(E) - L(E) + \frac{\partial}{\partial E}\left[J(E)\right] . \quad (21)$$

where $N(E)$ is the equilibrium-proton-density energy spectrum, $S_p(E)$ is the source of protons, $L(E)$ is the loss term, and $J(E)$ is the "energy current" $= N(E)\ dE/dt$ for equilibrium $dN(E)/dt = 0$. Let us now consider two special cases of this equation for equilibrium.

Case A: $L(E) = 0$.

For protons between 5 and 100 Mev the dominant loss process is slowing down by exciting and ionizing electrons by distant coulomb collisions, and we can ignore other losses. This slowing down contributes to the energy-current term and is not considered here to be part of $L(E)$; equation 21 becomes now

$$S_p(E) = \frac{\partial}{\partial E}\left[N(E)\frac{dE}{dt}\right] . \quad (22)$$

The proton-source term is given by the neutron's decay density from equation 17:

$$S_p(E) = \frac{dn(E)}{dV} = \frac{\eta\, 0.8\, E^{-2.0}}{\gamma v \tau_n}\left(\frac{R_e}{R_e + h}\right)^3 . \quad (23)$$

This expression is valid at low latitudes and close to the earth. The coefficient η is put in here because not all of the neutrons that decay form protons that are trapped.[38] Some of the protons made by neutron decay have pitch angles that are so small that they will hit the earth before they mirror. These protons will not form part of the trapped radiation. The coefficient η is called an injection coefficient and gives the fraction of protons that are trapped. The average value of η for inner-belt protons is $\bar{\eta} \approx 0.30$.[39]

We can solve equation 22 approximately[33,40] for the energy range 10 to 80 Mev by writing $dE/dt = v\ dE/dx$ and approximating $dE/dx = 243\rho E^{-0.79}$ Mev/cm; $v = 1.45 \times 10^9 \times E^{0.477}$ cm/sec; and $\gamma = 0.93E^{0.32}$, where ρ is the air density in gm/cm^3. Substituting in equation 22 we get

$$\frac{(0.30)\left(0.8E^{-2.0}\right)\left(\frac{6400}{7500}\right)^3}{\left(0.93E^{.032} \times 1.45 \times 10^9 E^{0.477} \times 1000\right)} = \frac{\partial}{\partial E}\left[kE^{-n}\, 243\,\rho\, E^{-0.79} \times 1.45 \times 10^9 E^{0.477}\right] . \quad (24)$$

Solving this gives

$$\Phi(E) = \nu N(E) = kE^{-n} = \frac{2.9 \times 10^{-16}}{\rho} E^{-0.72} \quad \text{for } 10 \text{ Mev} < E < 80 \text{ Mev}. \quad (25)$$

If a time-averaged density of $\rho = 2.8 \times 10^{-18}$ gm/cm^3—corresponding to an atomic density of 1.0×10^5 atoms/cm^3—is used, then $(E) = 110E^{-0.72}$ protons/cm^2-sec-Mev. This density is quite reasonable for this situation. The diurnal average density at 1100 km near solar maximum is about 2.5×10^5 atoms/cm^3.[41] The time-average density along a particle orbit is less than this by about a factor 2,[42] so the density used is quite a good one. This expression for $N(E)$ fits the Freden and White data well.

The lifetime τ_p of these protons can be obtained by using the "leaking bucket" equation.[12] We have for this problem, for equilibrium,

$$\text{Input} = \text{Output} = \frac{\text{Contents}}{\tau_p} , \quad (26)$$

which gives

$$\tau_p = \frac{\text{Contents}}{\text{Input}} = \frac{N(E)}{S(E)} .$$

We know $S(E)$ from equation 23 and $N(E)$

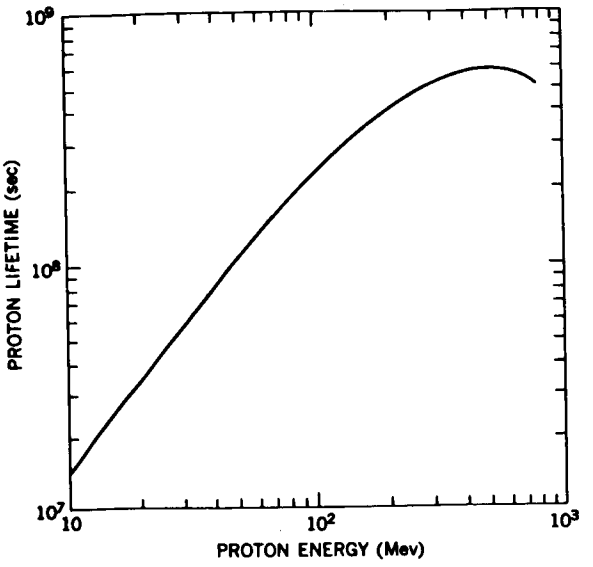


FIGURE 6.—Proton lifetimes at 1100 km in the inner belt based on slowing down and nuclear collisions.

from the solution of equation 24 (shown as the solid curve on figure 5); we can therefore calculate the lifetime directly. This gives $\tau_p = 7.0 \times 10^5 E^{1.3}$ sec for 10 Mev $< E <$ 80 Mev for 1100 km altitude. Figure 6 shows values of τ_p. For other oxygen densities the lifetime τ_p varies inversely as the time average density.

Case B: $de/dx \approx 0$.

If some other loss process occurs considerably faster than slowing down, then the protons will have essentially constant energy; slowing down can then be neglected. This situation is approximated for protons of $E > 300$ Mev. For these energies, the protons almost all have nuclear collisions before they slow down. The cross section σ for an inelastic collision of a high-energy proton with an oxygen nuclei is 3.0×10^{-25} cm^2. This gives a mean free path for nuclear interactions of $\lambda = \rho/n\sigma = 2.67 \times 10^{-23}$ gm/atom/3.6×10^{-25} cm^2/atom $= 74$ gm/cm^2. The range of a 500 Mev proton in oxygen is 135 gm/cm^2; therefore, these high-energy protons will usually have nuclear collisions before slowing down much. In this case, equation 21 becomes

$$S(E) = L(E) \tag{27}$$

For the loss term here we can write

$$L(E) = N(E)\, n\sigma\nu \ , \tag{28}$$

where n is the atomic density in atoms/cm^3; this gives

$$N(E) = \frac{\eta\, 0.8\, E^{-2.0} \left(\frac{R_e}{R}\right)^3}{\gamma \nu \tau_n} \left(\frac{1}{n\sigma\nu}\right) . \tag{29}$$

When we substitute

$$\nu = 2.69 \times 10^9\, E^{0.344} \ ,$$

$$\gamma = 0.428\, E^{0.205} \ ,$$

$$n = 1.0 \times 10^5 \text{ atoms/cm}^3 \ ,$$

$$\sigma = 0.36 \times 10^{24} \text{ cm}^2 \ ,$$

which are valid for 80 Mev $< E <$ 700 Mev, equation 29 becomes

$$\Phi(E) = \nu N(E) = 4.2 \times 10^6\, E^{-2.54} \frac{\text{protons}}{\text{cm}^2\text{-sec-Mev}} . \tag{30}$$

This is an asymptotic expression for $N(E)$ at high energies where slowing down is not important. Above 300 Mev it holds quite well but below this slowing down also is important.

The general form of equation 21 has been solved[33] to give the solid curve in figure 5. The form of the solid curve above 300 Mev agrees well with equation 30, and below 80 Mev the curve is exactly given by equation 25. There is no arbitrary normalization involved here; there are no adjustable parameters in the theory—the comparison of experiment and calculation is

direct. We know the source strength and properties, we know the loss processes and rate, and we know the properties of the atmosphere: from these data we get directly the proton-energy spectrum. The agreement with the data here is so good that the analysis is quite certainly correct, and neutron decay is the source of these protons.

Spatial Distribution

The spatial distribution of the inner-belt protons was first measured by the GM counter in Explorer IV.[10] When these data were obtained it was not known that the counting rate of this

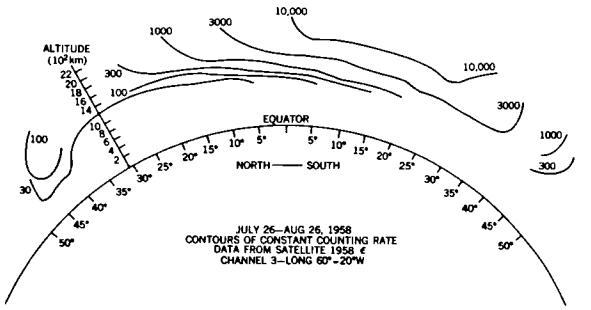


FIGURE 7.—A meridian section through the earth showing observed contours of the count rate of the unshielded GM tube on the Explorer IV satellite for the period July 26 to August 26, 1958, within longitude range 80°±20° W. Reproduced with permission from reference 10.

detector was due to protons, but it is now quite certain that this is the case. The 302 GM counter on board counted protons of $E_p > 30$ Mev and electrons of $E_e > 3$ Mev. The electron flux above this 3 Mev energy limit in the inner belt is thought to be quite small.

Contours of constant count rate measured on Explorer IV by Van Allen are shown in figure 7. The count rates increase with altitude, and show a fairly complicated change with latitude and longitude. Figure 8 gives curves (taken from reference 43) of the increase of count rate with altitude at different locations. The data curves O_1 and O_2 were obtained near Singapore; E_1, E_2,

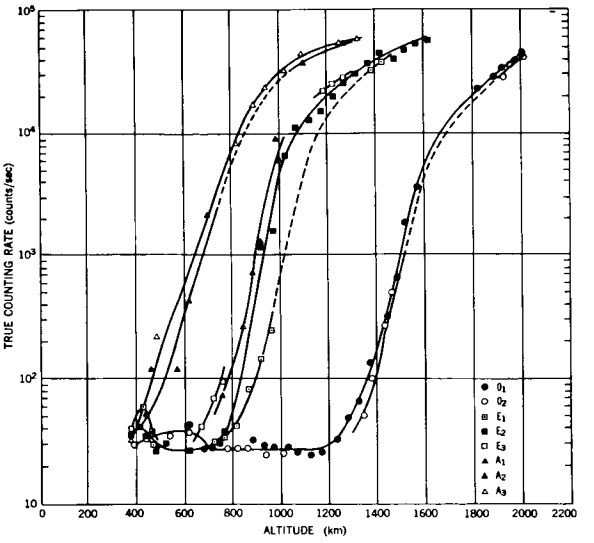


FIGURE 8.—Counting rate as a function of altitude for the GM counter on the Explorer I satellite for several different positions near the magnetic dip equator. The curves labeled "O" are data near Singapore, those labeled "E" are near Nigeria, the A_1 curve is in northern South America and A_2 and A_3 are in central South America. Reproduced with permission from reference 43.

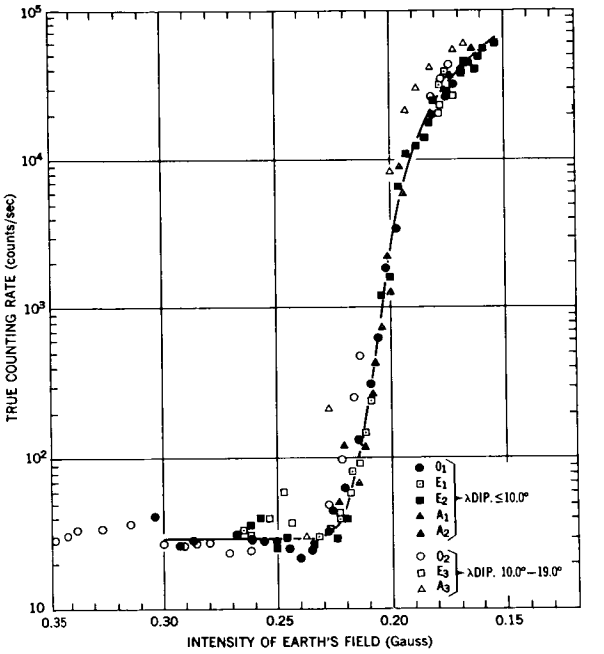


FIGURE 9.—The same counting rate data as shown in figure 8 replotted in terms of computed magnetic field strength at the observing point. All of the curves from figure 8 are now essentially superimposed. Reproduced with permission from reference 43.

and E_3 near Nigeria, Africa; A_1 in northern South America; and A_2 and A_3 in central South America. All of these locations are at about the same magnetic latitude, but the curves show quite different altitude behaviors. The reason for this is that the earth's surface magnetic field is different at these different locations. When these data are replotted in terms of the magnetic field instead of altitude at each location (figure 9) the curves then become essentially indistinguishable.[43] The reason for this is easily understood in terms of the motion of charged particles in a dipole magnetic field. The particles bounce back and forth, mirroring at one particular value of the magnetic field B. The particles also drift around the earth staying on a surface of constant integral invariant I; this corresponds to staying at a constant magnetic latitude. Therefore, for observations at one latitude the count rate should vary with B as it does in figure 8. The altitude of the mirror points represented by the count rates in figure 8 get as low as 400 km over South America. Below this altitude, galactic cosmic rays provide most of the count rate. The lower edge of the trapped radiation belt is clearly controlled by the atmosphere. Protons are lost by slowing down. A 50 Mev proton, to be brought to rest, must transit about 2.5 gm/cm^2 of oxygen. At 400 km, the atmospheric density is about 5×10 atoms/cm^3 or about 10^{-14} gm/cm. At velocity of 10^{10} cm/sec, the proton will take about 2×10^4 sec to slow to a stop. But, because of the variations or atmospheric density along the particle's orbit—especially the variation due to the drift in longitude—the lifetime is increased probably to 10^6 sec. The source strength of fast protons from neutron decay near the earth is about 10^{-15} protons/cm^3-sec but perhaps only 30 percent of these will be going in the proper direction to be trapped. We get an equilibrium flux here of

$$\Phi_p = \left(0.3 \times 10^{-15} \frac{\text{protons}}{\text{cm}^3\text{-sec}}\right)\left(10^{10}\frac{\text{cm}}{\text{sec}}\right)(10^6 \text{ sec}) = 0.3\frac{\text{protons}}{\text{cm}^2\text{-sec}} .$$

This calculation is not very exact, but the value of the flux thus found is comparable with the cosmic-ray flux of about 2 protons/cm^2-sec; therefore, it is reasonable that below this altitude of 400 km the trapped flux gets too small to be measured and cosmic rays dominate.

We are quite certain the lower edge of the inner belt of high-energy protons is controlled by the earth's atmosphere. But we are not nearly so certain what controls the outer edge of the proton belt. We would expect, from what is known of the inner belt, that it should extend out many earth radii. The neutron decay source strength decreases about as $1/R^2$, but the proton lifetime for slowing down should go up about as fast as the source strength goes down, so the equilibrium proton flux expected in the outer belt is about the same as that observed in the inner belt. Actually, the observed trapped high-energy proton flux in the outer belt is less than the cosmic-ray flux and consistent with zero. Fan, Meyer and Simpson[44] gave an outer-belt proton flux of

$$\Phi_p = 0 \pm 0.1 \frac{\text{protons}}{\text{cm}^2\text{-sec}} \text{ for } E_p > 75 \text{ Mev}.$$

This means that the lifetime of the protons in this region is reduced by a factor of 10^4 or more by some additional process.

The processes which seem most probable to cause this reduction in lifetime have to do with time variations of the magnetic field. Welch and Whitaker[45] suggested that time or space variations in the magnetic field could produce "magnetic scattering" of trapped particles. If the magnetic perturbations are of such character that they cause a breakdown of the adiabatic invariants then the particle's motion will be altered. Hydromagnetic (hm) waves having a wave length $\lambda < R_c$, where R_c is the particle cyclotron radius, will break down the magnetic moment invariant. When the particle encounters the wave a change in μ will take place, and as the result the particle's

mirror point will be changed. Repeated encounters with waves will cause a diffusion of the particle's mirror point and result in a loss of particles out of the loss cones into the atmosphere.

Dragt[46] has calculated the effect of hydromagnetic waves on a high-energy proton magnetic moment. He finds that for *hm* wave frequencies of a few cps and for particles traveling along field lines, a condition can exist such that the *hm* wave frequency is Doppler shifted to equal the particle's cyclotron frequency. For this resonant condition the particle's magnetic moment is changed as a result of interacting with the wave.

The particle lifetime τ against *hm* wave scattering is given by Dragt as

$$\tau = \frac{2\tau_B}{\pi^2 p^2}\left(\frac{\Delta B_{hm}}{B}\right)^2 , \qquad (31)$$

where τ_B is the bounce period and p is the number of *hm* waves encountered per bounce. For 100 Mev protons at $2R_e$ the bounce period $\tau \approx 0.1$ sec, and $B=0.04$ gauss. Assuming $\Delta B=3$ gamma and $p=1$ we get

$$\tau \approx \frac{1}{2} \text{ day}.$$

This lifetime is short enough that the equilibrium flux of protons in the outer belt would not be measured because the cosmic-ray flux is larger.

Also, the radial dependence of the scattering is reasonable. At $2R_e$, the maximum energy trapped is proportional to R^{-11}. But at smaller altitudes, the scattering is less. The count rate versus altitude curve of the GM counter on Pioneer III is reasonably well reproduced by this analysis. The requirements on *hm* wave characteristics here are quite reasonable. A wave characteristic of $\Delta B=3\gamma$ at 3 cps is not a very large disturbance.

Wentzel[47] made a similar analysis of the breakdown of the magnetic moment of high-energy protons by *hm* waves. He found that the magnetic moment was changed for encounters with *hm* waves when $v/v_c>0.4$, where $v_c=hm$ wavelength/larmor period. Accordingly, 56 Mev protons should extend out to 6000 km and 28 Mev protons out to 7000 km. These results are relatively similar to Dragt's results.

Parker[48] has considered how a breakdown of the integral invariant I can transport and accelerate particles. If the magnetic field at the mirror point changes in a time short compared to a bounce period, then I is not a constant of the motion. The bounce period is[49]

$$\tau_B \approx \frac{4r_0}{v} \approx \frac{4\times 10^9 \text{ cm}}{10^{10}\ \frac{\text{cm}}{\text{sec}}} = 0.4 \text{ sec}. \qquad (32)$$

So, waves having a frequency of 5 or 10 cycles/sec will cause nonadiabaticity here. These waves at the mirror point result in a Fermi acceleration.[50] Some particles encounter mirror points moving towards them due to the field perturbation ΔB and some particles find mirror points moving away from them. But, because statistically more approaching collisions take place, there is a net energy gain given by

$$\frac{dE_{\|}}{dt} = 8E_{\|}\left(\frac{v}{\omega_{\|}}\right)^2 p , \qquad (33)$$

where $E_{\|}$ and $\omega_{\|}$ are the particle energy and velocity component along the field line, respectively, p is the number of mirror point reflections per unit time, and v is the velocity of the mirror point motion.

The interesting feature here is that as the particle gains energy its mirror point is systematically lowered because all the energy gain is in $E_{\|}$ and, therefore, the pitch angle decreases and the particle is lost into the atmosphere. For a $\Delta B/B$ of 10^{-3} and constantly moving mirror points, the characteristic time for a particle to diffuse down into the atmosphere is about 3×10^6 sec. It is not known whether these hydromagnetic wave conditions are really met, and thus whether this lifetime is reasonable or not. Parker made this analysis for outer-belt 100 kev electrons, but it applies equally to protons of about 50 Mev.

A proton lifetime of 3×10^6 seconds and a source strength of about 0.2×10^{-15} protons/cm^3-sec gives an equilibrium flux of

$$\Phi_p = \left(0.2\times 10^{-15}\ \frac{\text{protons}}{\text{cm}^2\text{-sec}}\right)\left(10^{10}\ \frac{\text{cm}}{\text{sec}}\right)\left(3\times 10^6 \text{ sec}\right) = 6\ \frac{\text{protons}}{\text{cm}^2\text{-sec}} .$$

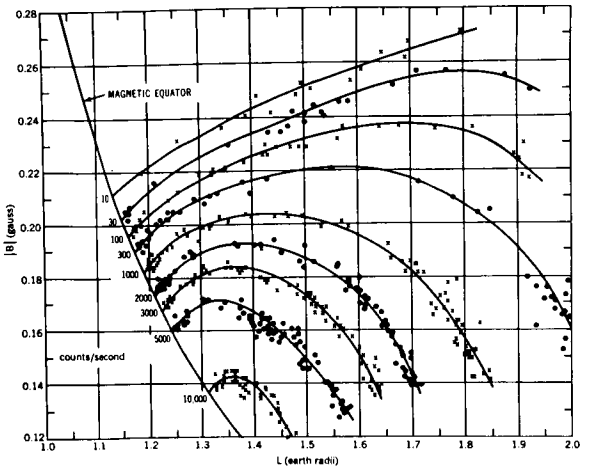


FIGURE 10.—Contours of constant true counting rate of the unshielded GM counter in Explorer IV. The points shown correspond to data obtained over a wide range of geographic latitude and longitude. Reproduced with permission from reference 51.

But this is larger than the measured outer-belt flux by at least an order of magnitude, so the lifetime must be less than 10^6 sec by a factor of 10 or more. From this analysis it seems that the breakdown of the integral invariant probably is not the dominant loss process, but that another process gets rid of the protons faster.

It would appear from the work of Dragt[46] and Wentzel[47] that the breakdown of the magnetic moment of the protons by *hm* waves caused the loss of protons in the outer belt. The *hm* waves required to do this seem available and the results of the calculations agree reasonably well with observations. We cannot be sure that *hm* wave scattering is the controlling factor for the outer edge of the inner belt because there is no direct verification of this, but it seems the best guess now.

If the outer edge of the proton belt is controlled by *hm* wave scattering as we think, then there may be an interesting solar cycle change in the proton flux at $2R_e$ and beyond.[46] If, at solar minimum, there is less energy in *hm* waves (as seems reasonable), then the proton lifetime will increase and an appreciable flux of protons may build up in the outer belt.

In order to organize the kind of data shown in figure 7 into some easily manageable form, McIlwain developed the B–L coordinate system[51] where B is the scalar magnetic field and L is a distance which, in a dipole field, is the distance to the equator for a particular field line. It is defined in terms of the integral invariant I in such a way that the real earth's field is used. L is very nearly constant along a field line, so it can replace the commonly used equatorial radius R_0, but use real values of the earth's field. The Explorer IV data in figure 7, when plotted in terms of B–L coordinates, can be combined for different geographic latitudes, longitudes, and altitudes into the simple form shown in figure 10. Reading down a line of constant L here corresponds to going out from the earth along a field line. This information can be transformed into a more familiar form by using

$$R = L\cos^2\lambda ,$$

and

$$B = \frac{M}{R^3}\sqrt{4 - \frac{3R}{L}} .$$

This R, λ presentation of the data in figure 11 has the earth's magnetic field transformed into a dipole field, but the earth's surface now has an odd shape.

Figure 11 is a representation of the fluxes of high-energy protons (E_p>30 Mev) in the inner belt as of 1958. It is probably accurate to within a factor 2 over most of the range of values. The

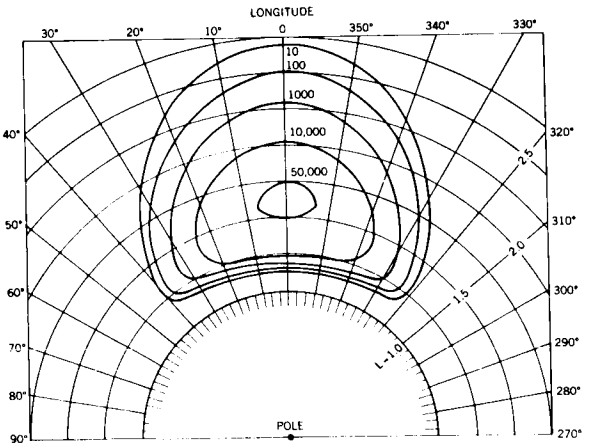


FIGURE 11.—High-energy-proton fluxes in the inner belt. This is a R–λ representation based on McIlwain's L parameter, On this picture the earth's surface is bumpy, but centered on $L=1$. The fluxes are given in protons/cm²-sec.

picture was somewhat different in 1961; this variation is discussed later.

In figure 11, the data from Explorer IV have been combined with data from Pioneer III to extend the range of B–L values covered. This, with some extrapolations, covers the whole inner belt. The count rates in figure 10 have been converted to particle fluxes in figure 11 by the use of the proper detector geometrical factors.

Pioneer III went out radially about 100,000 km from the earth and then returned.[12] It cut through the inner belt twice. The 302 GM counter on board was nearly identical in threshold to the counter in Explorer IV; however, it had a slightly larger geometrical factor $G_0=0.62$.[12] The count rate curve for the reentry flight is shown in figure 12. The section from 15,000 km inwards is the important region for us. This shows the nature of the outer edge of the proton belt within about 20 degrees of the equator and, when combined with Explorer IV data, allows us to make a more complete map of the inner-belt-proton flux.

Time Variations

The protons in the inner radiation belt show relatively small changes in intensity with time in contrast to the outer belt, which varies considerably, especially at the time of magnetic storms.

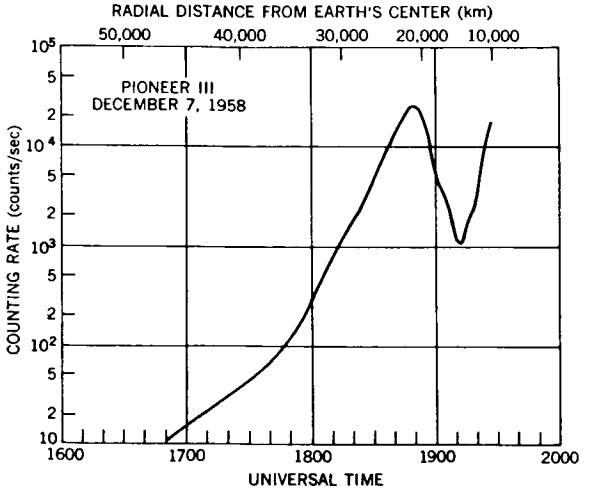


FIGURE 12.—True count rate of the 302 GM counter on Pioneer III on the inbound flight. Reproduced with permission from reference 12.

Changes in the proton flux at low altitudes were not seen even in connection with magnetic storms on Explorer IV.[52] There may be some small time changes in the proton energy spectrum: a small peak at about 75 Mev seemed to be present in the proton energy spectrum measured in July 1959;[37] then in October 1960 a peak was observed at 30 Mev.[32] There is a possibility that it is the same peak degraded in energy, and that it was the result of neutrons produced in one polar cap proton event.[53] There is little doubt that polar cap protons are important in generating neutrons, but it seems questionable whether this structure in the trapped proton energy spectrum is the result of polar cap protons. It now appears that if there are changes in the spectral shape above 30 Mev, they are quite small.[32]

Time changes in the inner belt have been detected by the 302 GM counter on Explorer VII.[55] It counted protons of $E_p>10$ Mev and electrons of $E_e>1.1$ Mev. At the outer edge of the inner belt for $1.8<L<2.2$ time variations of a factor of 3 or more are seen at the times of magnetic storms. The changes may well be in the electron population. For $1.5<L<1.8$ there are small variations seen in the count rate at storm times, maybe in the proton population or maybe not.

From November 1959 to November 1960 a rather gradual increase of a factor of 2 in intensity was seen in the inner part of the inner belt ($L<1.5$) by Explorer VII. At least part of this increase is expected, on two grounds: (1) the cosmic-ray flux increases by about a factor of 2 from solar maximum to solar minimum, and (2) the exospheric densities will decrease. The heating of the exosphere is less at solar minimum, so that the scale height decreases and therefore the densities decrease.

Recent calculations on the solar cycle change of exospheric density[41] give the atomic densities listed in Table 3. Both the slowing down loss process and the nuclear collision loss process will vary with these two densities. The slowing down rate dE/dx goes as

$$\frac{dE}{dx} \propto [4\,N(O) + N(He)] \cdot$$

TABLE 3.—*Atomic Densities at Various Altitudes*

Altitude (km)	Atomic Densities (atoms/cm^3)			
	Near Solar Maximum†		Near Solar Minimum††	
	N(O)	N(He)	N(O)	N(He)
1000	7×10^5	4.4×10^5	440	1.1×10^5
1500	1.1×10^4	1.6×10^5	18	1.6×10^4
2000	290	6.2×10^4	0	2800

†S=200 from Reference 41.
‡S=70 from Reference 41.

Oxygen is more important here because it has 4 times as many electrons to help in slowing down. The nuclear collision rate C shows a similar variation

$$C \propto [2.5\,N(O) + N(He)]$$

The oxygen nucleus has a large cross section σ for nuclear collisions ($\sigma \propto A^{2/3}$), so oxygen has more collisions per nucleus than helium does.

From figure 6 we got the proton lifetime at 1000 km altitude for near solar maximum given in Table 4. The other values in the table are obtained from these 1000 km values by using density variations in Table 3 and equation 26. The period of the solar cycle $\tau_{sc}=4\times10^8$ seconds. If the proton lifetime is short compared to this, then the particle population will change with changes in source strength or loss rate; but if the proton lifetime is long compared to τ_{sc}, then the solar cycle will be averaged out and no population changes will occur. We see from Table 4 that changes in low-energy particles at 1000 km will be appreciable, but at 2000 km no changes will occur. The factor of 2 increase in the number of protons of $E_p>19$ Mev seen by Explorer VII at 1000 km during 1960[54] is consistent with the increase expected here on the basis of solar cycle changes in cosmic ray flux and exospheric density.

MEDIUM-ENERGY PROTONS

Out to about $L=1.6$, the proton energy spectrum is as shown in figure 5; but for $L>1.7$, Naugle and Kniffen[35] found that for $E_p<30$ Mev the spectrum showed a sharp rise above this curve. Apparently some additional process comes into play to produce these protons of $E_p<30$ Mev near the outer edge of the inner belt. The Naugle and Kniffen data are given in figure 13.

A second experiment apparently saw this same group of particles: A proton spectrometer flown on a Scout vehicle to an altitude of 4800 km found a large number of protons[55] of 1 Mev $<E_p<10$ Mev. One part of Naugle and Kniffen's data was

TABLE 4.—*Inner Belt Proton Lifetimes*

Energy (Mev)	Lifetime					
	Near Solar Maximum (sec)			At Solar Minimum (sec)		
	1000 km	1500 km	2000 km	1000 km	1500 km	2000 km
15	2.4×10^7	3.8×10^8	1.4×10^9	7.2×10^8	4.8×10^9	2.4×10^{10}
40	8.5×10^7	1.2×10^9	5.2×10^9	2.5×10^9	1.7×10^{10}	8.4×10^{10}
100	2.8×10^8	4.2×10^9	2.0×10^{10}	8.4×10^9	5.6×10^{10}	2.8×10^{11}

at 1884 km altitude at $L=1.722$ and could be fit well by

$$\Phi_p(E) = 6.8 \times 10^6 E^{-4.5} \frac{\text{protons}}{\text{cm}^2\text{-sec-ster-Mev}} \text{ for } 10 \text{ Mev} < E_p < 50 \text{ Mev.}$$

Bame, et al.[55] made measurements at 4800 km at $L=2.50$ and got a proton flux given by

$$\Phi_p(E) = 2 \times 10^6 E^{-5.2} \text{ for } 1 \text{ Mev} < E_p < 2.24 \text{ Mev}$$

and

$$\Phi_p(E) = 0.71 \times 10^6 E^{-3.9}$$

$$\text{for } 2.2 \text{ Mev} < E_p < 7.3 \text{ Mev.}$$

Bame, et al. showed that an extrapolation of the Naugle–Kniffen data down to 1 Mev gave moderately good agreement with their data (see figure 14). Therefore, it appears that these two experiments were observing the same type of particles.

What new source of particles can produce medium-energy trapped protons only above $L=1.7$? Armstrong, et al.[34] suggested that solar

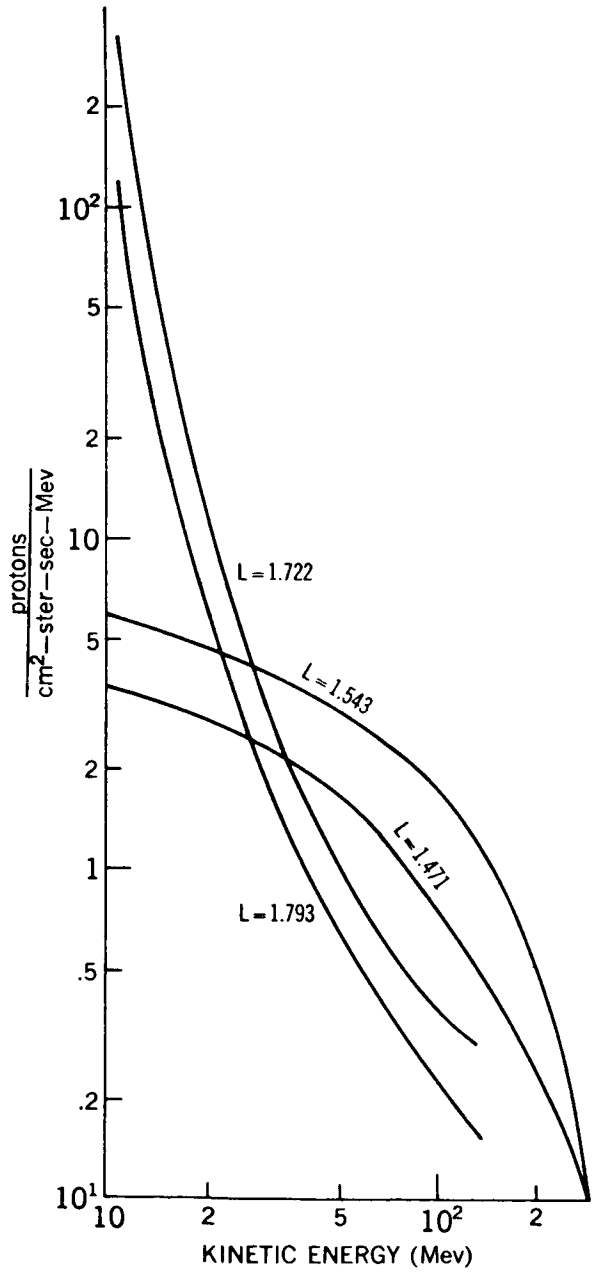


FIGURE 13.—Proton-energy spectra measured by Naugle and Kniffen for different positions along the vehicle trajectory. Reproduced with permission from reference 35.

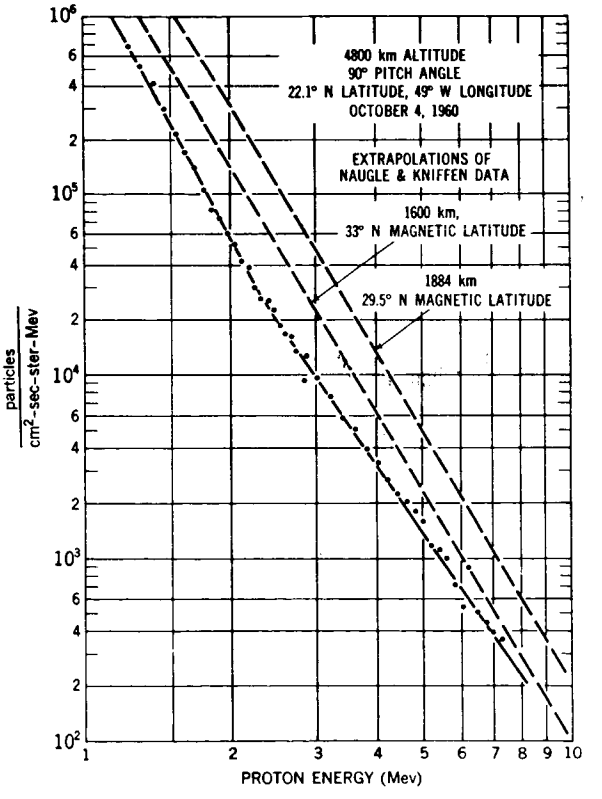


FIGURE 14.—Energy spectrum at 4800 km for protons measured by Bame, Conner, et al. Two of the spectra obtained by Naugle and Kniffen have been extrapolated to lower energies and are shown. Reproduced with permission from reference 55.

protons bombarding the polar cap at the time of solar flares should produce neutrons, and that the decay of these neutrons then constitutes a new source of protons. We are sure that galactic cosmic ray protons produce neutrons which then produce inner belt protons; so it seems very reasonable that solar protons should similarly produce neutrons and thereby trapped protons.

The trapped protons produced from solar protons will be different from the other trapped protons in two ways. First, their spatial distribution will be different. This is because solar protons arrive at the earth only in the polar regions (because of their low energy, the earth's field prevents them from getting further from the poles than about 55 degrees of magnetic latitude). The inner part of the inner belt is therefore geometrically shadowed from this source, so more of these neutrons can produce trapped protons near the inside of the inner belt. Lenchek[56] has calculated that no trapped protons will be produced for $L < 1.65$. This is consistent with the data of Naugle and Kniffen.[35]

The second distinctive feature of trapped protons from solar protons is that the intermediate neutrons produced this way have lower average energies than those from the galactic cosmic ray source. The solar proton energy spectrum usually extends up past 100 Mev, but most of the protons have much lower energies. Therefore, the neutron energy spectrum will not contain many particles of $E > 50$ Mev and the peak of the neutron spectrum should be at about 1 Mev. Lenchek[56] approximated this neutron spectrum $S_n(E)$ by an evaporation spectrum

$$S_n(E) = kEe^{-E/A},$$

and evaluated the total flux of trapped protons made by this process.

A 10 Mev proton that mirrors at 1500 km has a lifetime for slowing down of about 10 years, so this process will average over a large number of solar events. Lenchek estimated the time-average solar-proton flux from data of Webber[57] to be about 10 protons/cm^2-sec-ster, which is much higher than the galactic cosmic ray flux. Using an average atmospheric density of 10^3 atoms/cm^3, he obtained the trapped-proton flux observed by Naugle and Kniffen.[35] These numbers seem a little extreme, but the idea seems completely reasonable.

It appears quite reasonable from this analysis that these low-energy protons are made from solar protons striking the polar atmosphere. The fact that there is a sharp inner edge on this type of trapped radiation strongly rules against direct solar injection or local acceleration.[56]

LOW-ENERGY PROTONS

One of the most interesting recent radiation-belt discoveries is the large flux of low-energy protons at 1000 km.[58] On Injun, a CdS detector measured a heavy ion energy flux of about 30 ergs/cm^2-sec-ster at 1000 km in the energy range of 0.5 kev to 1 mev. These particles are probably protons. If they have an average energy of 100 kev, there is a proton flux of

$$\Phi_p = \frac{\left(50 \frac{\text{ergs}}{\text{cm}^2\text{-sec-ster}}\right)(1 \text{ ster})}{\left(0.1 \frac{\text{Mev}}{\text{proton}}\right)\left(1.6 \times 10^8 \frac{\text{erg}}{\text{Mev}}\right)} = 3 \times 10^8 \frac{\text{protons}}{\text{cm}^2\text{-sec}} .$$

This is a quite large flux. The spatial distribution of these low-energy protons is very similar to the distribution of high-energy inner-belt protons.

This strong similarity might be explained by: (1) the two groups of particles coming from a common source, (2) the high-energy protons generating the low-energy protons. We are very sure the high-energy protons are made by neutron decay. Also, slow protons produced by neutron decay do exist. We can evaluate the possibility that the observed slow proton flux results from neutron decay. Most of the neutrons in the leakage spectrum are between 100 kev and 10 Mev. Let us say that one half of the neutron-decay events at 100 km produce protons of about 1 Mev. From figure 3 we see that about 2×10^{-12} protons/cm^3-sec will be made in the Mev range from these decay events. If these protons

are lost by slowing down, they must encounter 3×10^{-3} gm/cm^2. In an atmosphere of $n=2.5\times10^5$ O atoms/cm^3 of oxygen (or $\rho=6\times10^{-18}$ gm/cm^3), and traveling at an average velocity of 5×10^8 cm/sec during the slowing down process, a 1 Mev proton will slow down in a time τ given by

$$\left(6\times10^{-18}\frac{\text{gm}}{\text{cm}^3}\right)\left(5\times10^{8}\frac{\text{cm}}{\text{sec}}\right)\tau = 3\times10^{-3}\frac{\text{gm}}{\text{cm}^3}$$

or

$$\tau = 10^6 \text{ seconds} .$$

We get from this the equilibrium-proton flux of

$$\Phi_p = \left(2\times10^{-12}\frac{\text{protons}}{\text{cm}^3\text{-sec}}\right)\left(5\times10^{8}\frac{\text{cm}}{\text{sec}}\right)(10^6 \text{ sec}) = 10^3\frac{\text{protons}}{\text{cm}^2\text{-sec}} .$$

This is much less than the observed flux of $\Phi_p\approx10^8$ so this process can not produce the observed protons.

Next, we can consider the possibility that the high-energy protons produce the low-energy protons. This might be done by coulomb collisions with the thermal protons. We can estimate the flux produced in this way. The cross section for coulomb collisions is

$$d\sigma = e^4\frac{2\pi\sin\theta\cos\theta\,d\theta}{E^2}\left(\frac{1}{\sin^4\theta}\right), \tag{34}$$

where E is the energy of the incident particle and θ is the scattering angle. For a range of values of θ, the struck proton will have energies E_s in the 0.5 kev to 1 Mev energy range given by

$$E_s = E\sin^2\theta . \tag{35}$$

Integrating over θ to get the total cross section and rearranging the coefficient gives

$$\sigma = \frac{0.5\times10^{-24}\text{ cm}^2}{\left(\frac{E}{mc^2}\right)^2}\int\frac{d\sin\theta}{\sin^3\theta} = \frac{0.25\times10^{-24}}{\left(\frac{E}{\text{cm}^2}\right)^2}\left[\frac{-1}{\sin^2\theta}\right]_{\theta^{min}}^{\theta^{max}} \tag{36}$$

for a 50 Mev incident proton. In order to get the range of E_s energies indicated, we need

$$\sin^2\theta^{max} = \frac{1}{50} \quad \text{and} \quad \sin^2\theta^{min} = \frac{.0005}{50} .$$

This gives a total cross section of $\sigma=2\times10^{-24}$ cm^2. Using a high-energy-proton flux of 10^4 proton/cm^2-sec and a thermal proton population of 10^4 atoms/cm^3, we get a slow proton product in rate S_p of

$$S_p = \left(10^4\frac{\text{fast protons}}{\text{cm}^2\text{-sec}}\right)\left(10^4\frac{\text{H atoms}}{\text{cm}^3}\right)(2\times10^{-24}\text{ cm}^2)$$

$$S_p = 2\times10^{-16}\frac{\text{protons}}{\text{cm}^3\text{-sec}} \text{ of } 0.5 \text{ kev} < E_p < 1 \text{ Mev.}$$

This source strength is considerably less than neutron-decay source strength, and the proton lifetimes will be shorter here too because the protons made are of lower average energy. Therefore, this source produces an even smaller fraction of the observed protons and is not important.

We can ask another question. Is there enough energy in the high-energy protons to supply the low-

energy protons by any process? Using a proton lifetime of 10^6 sec and velocity of 5×10^8 cm/sec, we get the energy requirement of the low-energy-proton source of

$$\left(50\ \frac{\text{ergs}}{\text{cm}^2\text{-sec}}\right)\left(\frac{1}{5\times10^8\ \frac{\text{cm}}{\text{sec}}}\right)\left(\frac{1}{10^8\ \text{sec}}\right) = 10^{-13}\ \frac{\text{ergs}}{\text{cm}^3\text{-sec}}\ .$$

The energy loss rate E_L of inner-belt high-energy protons is about

$$E_L = \left(50\ \frac{\text{Mev}}{\text{protons}}\right)\left(1.6\times10^{-6}\ \frac{\text{erg}}{\text{Mev}}\right)\left(10^4\ \frac{\text{protons}}{\text{cm}^2\text{-sec}}\right)\left(\frac{1}{10^{10}\ \frac{\text{cm}}{\text{sec}}}\right)\left(\frac{1}{10^8\ \text{sec}}\right) = 8\times10^{-19}\ \frac{\text{erg}}{\text{cm}^3\text{-sec}}\ ,$$

where we have assumed an average lifetime of 10 seconds.

We see that energetically the high-energy protons cannot produce the low-energy protons by any process. The energy in the low-energy protons must come from some other sources that are currently not understood. These low-energy protons are one of the most interesting mysteries of the radiation belt.

ELECTRONS

It seems surprising, but we know a good deal less about the electrons in the inner belt than about those in the outer belt. One reason for this is that the high-energy protons present in the inner belt tend to hide the electrons. The protons are counted with high efficiency by most detectors and penetrate considerable amounts of shielding. Instruments designed to count electrons will usually count the protons too.

There are, however, several experiments that have given some information about the inner belt electrons.

Measurement of the Electron Spectrum

Several particle spectrometers have been flown on pods on Atlas rockets, by Holly, et al.[59] to altitudes up to 1500 km in the inner belt near the equator. Electrons were identified by the use of magnetic analysis. The differential energy spectrum of the electrons measured here is shown in figure 15.

The inner belt electron spectrum has also been measured recently by a 10-channel magnetic spectrometer flown on Discoverer satellites at about 400 km.[60] In the Atlantic Ocean off Brazil, a region of high count rate was observed; this region connects to the inner belt.[61] Vernov found that this region had a high content of protons by comparing the count rates of a GM counter and a scintillation counter. The spectrometer on Discoverer satellites 29 and 31 had a blank channel to count protons which were considered a background for this experiment and subtracted from the data. The electron spectrum measured in the South Atlantic by a Discoverer is shown in figure 15.

The count rate here is due to inner belt electrons that have come down to low altitudes at this location because the magnetic field is weak. This spectrum extends to higher energies and is flatter than the Atlas pod spectrum. The difference in these two spectra is not understood; they were not measured at the same time or place, so a direct comparison is really not possible. The Discover spectrometer did see other

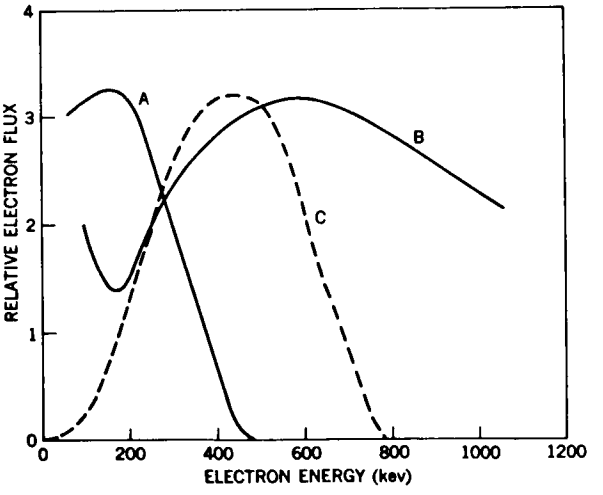


FIGURE 15.—Electron-energy spectra in the inner belt: Curve A measured by Holly, Allen, and Johnson; curve B measured by Bame, Conner, et al. (reference 55); curve C is a calculated spectrum based on a neutron decay source.

types of electron energy spectra in other places, but never any type resembling the spectrum of reference 59. In the South Atlantic the only spectrum seen is that shown in figure 15.

The energy spectrum expected[62,63,64] in the inner belt from neutron β-decay electrons is also shown in figure 15. Electrons are lost from the inner belt by coulomb collisions which change the pitch angle and cause the particles to diffuse out of the loss cone. The energy dependence of the electrons lifetime τ as calculated for scattering is[65]

$$\tau \propto v p^2 ,$$

where p = electron momentum and v = electron velocity. The shape of the equilibrium electron spectrum expected from neutron β-decay is simply the shape of the β-decay spectrum (see figure 2) weighted by the particles' lifetimes which change with energy as given above. This calculated spectrum lies between the two experimental spectra.

The situation here is not clear. It appears experimentally that the inner belt electron spectrum may vary with space or time. But the observed spectra do not seem to agree very well with the spectrum calculated from neutron decay. It appears that sometimes there are electrons of $E>1.2$ Mev in the inner belt. These cannot be directly made by neutron decay. Some acceleration process must act to produce these high-energy electrons. It would seem that neutron decay electrons could be accelerated most easily because they start with relatively high energies to begin with. We must wait for more experimental information to make any kind of quantitative comparison with calculations based on neutron decay.

The Electron Flux

There have been several measurements which give the electron flux in the radiation belts. None of these are for the central regions of the inner belt, but we can interpolate what measurements are available to give reasonable estimates in the inner belt.

There are three measurements of the electron flux at about 1000 km altitude. O'Brien, et al.[66] have measured the flux of electrons of $E_e>40$ kev at 1000 km on the Injun satellite by using magnetic spectrometers. The flux thus determined was about 10^5 to 10^6 electrons/cm^2-sec-ster so, because $\Delta\Omega\approx 1$ ster at these altitudes, the omnidirectional flux is $\Phi_e\approx 10^5$ to 10^6 electrons/cm^2-ster. These measurements were for $L>2$ so they are not really inner-zone electron fluxes. The electron flux of $E_e>30$ kev measured by Holly, et al.[59] at 980 km was about 7×10^5 electrons/cm^2-sec. This was at 15°N latitude and is truly an inner belt flux. Cladis, et al.[67] measured an electron flux for $E>50$ kev of 4×10^6 electrons/cm^2-sec at $L\sim 2.4$ using a magnetic spectrometer; these measurements give an electron flux at 100 km of the order of 10^6 electrons/cm^2-sec of $E>40$ kev.

Equipment on the Explorer XII (1961υ) satellite has measured the electron flux in the outer belt. O'Brien, et al.[68] report fluxes going up to 10^8 electrons/cm^2-sec of $E>40$ kev, but a more typical flux[69] is about 10^7 electrons/cm^2-sec. The electron flux is quite constant from about 50,000 km into about 10,000 km.

From this data we can guess that the electron flux in the center of the inner belt is about 10^7 electrons/cm^2-sec, possibly going as high as 10^8 electrons/cm^2-sec.

One early measurement seemed to give information about the inner-belt electron flux, but this now appears incorrect.

The Explorer IV satellite carried four detectors that detected particles of different ranges. The particles that penetrated 1 gm/cm^2 or more of shielding were identified as being high-energy protons.[13]

One detector on Explorer IV, the energy scintillator, counted particles that penetrated 1.0 mg/cm^2 of shielding. A large flux, up to 10 ergs/cm^2-sec-ster or more, was measured by this detector. A tentative analysis of this experiment[13] suggested it was due to electrons that would just penetrate the 1 mg/cm^2 foil of about 20 kev. This model was proposed by analogy with early ideas about the composition of the outer belt. Assuming the particles were 20 kev, it took about 2×10^9 electrons/cm^2-sec-ster to give the observed energy flux.

But it now appears that this interpretation was not correct. Recently, on the Injun satellite, an energy flux of up to 100 ergs/cm^2-sec of heavy ions was measured.[58] These were very probably protons and they were in the energy range of

0.5 kev to 1 Mev. Many of these particles would be counted by the Explorer IV energy scintillator. It, therefore, seems likely that this Explorer IV counter was counting low-energy protons and not electrons.

The inner-belt electron flux expected from neutron decay is of the same order of magnitude as the observed flux. Wentworth, et al.[65] have calculated the electron lifetime for coulomb scattering by solving the Fokker–Planck equation. For an atmospheric He density at 2000 km of $\bar{\rho}=10^5$ atoms/cm^3 they get a lifetime for a 300 kev electron of $\tau_e=1.5\times10^8$ sec. Using this lifetime and the source strength[26] of 3×10^{-12} electrons/cm^3-sec, we get an equilibrium electron flux of

$$\Phi_e = \left(3\times10^{12}\frac{\text{electrons}}{\text{cm}^3\text{-sec}}\right)\left(3\times10^{10}\frac{\text{cm}}{\text{sec}}\right)(1.5\times10^8\ \text{sec}) \approx 10^7\frac{\text{electrons}}{\text{cm}^2\text{-sec}}\ .$$

In comparing the observations of electrons with calculations based on the neutron-decay theory, we find that the calculated and observed fluxes are quite similar but that the spectra are not in as obvious agreement. It is quite certain that neutron-decay electrons produce a large fraction of the trapped electrons. It is also quite certain that acceleration processes also operate to generate the high-energy electrons.

OTHER PARTICLES

If the sun were the source of particles in the inner radiation belt we would expect to find not only protons but other heavier particles such as deuterons, tritons, and He^3 and He^4 nuclei. The sun contains about 15 percent He nuclei[70] and solar cosmic rays contain about 5 percent He nuclei.[71] The lifetime of a He^4 nuclei in the inner belt would be about five times less than that of a proton of the same energy because the rate of slowing down is faster for $Z=2$. On this basis, we would expect about 1 percent He^4 in the inner belt. But experimentally not one $Z=2$ track has been found in nuclear emulsion. The total nummer of particles measured in three experiments is given in Table 5. An upper limit of the α flux is 0.1 ± 0.05 percent of the proton flux in the energy interval 125 to 185 Mev.[36] This quite clearly shows that the sun contributes few, if any, of the heavy particles in the inner belt.

A few deuterons and tritons (roughly $\frac{1}{2}$ percent each) were found in the emulsion experiments as is shown in Table 5. These particle fluxes can be explained[33] as being the result of nuclear collisions of trapped protons with O and N nuclei in the very thin atmosphere present at radiation belt altitudes. No heavy particles have been observed in the inner belt that cannot be understood by the neutron-decay source.

CONCLUSIONS

From the foregoing discussion we can draw the following conclusions about the inner radiation belt:

The high-energy protons ($E_p>30$ Mev) result from galactic cosmic-ray production of neutrons

TABLE 5.—*Measurements on Trapped Heavy Particles*

Experiment	No. of protons	No. of deuterons	No. of tritons	No. of alphas
Freden and White, May 1959 (reference 33)	243	0	3	0
Armstrong, et al., July 1959 (reference 34)	477	5	0	0
Heckman and Armstrong, October 1960 (reference 36)	301	0	0	0
Total	1021	5	3	0

in the earth's atmosphere and the subsequent escape and decay of these neutrons. Both the energy spectrum and the intensity of the trapped protons are completely consistent with this interpretation. The time variations observed in this proton population are understood well in terms of solar-cycle variations in the exosphere density and of cosmic-ray flux.

The spatial distribution of these protons also seems reasonable on the basis of neutron decay injection. Protons are lost from the belt by slowing down. Increasing density of air limits the lower edge of the belt, and the changes in altitude of the lower edge with longitude are due to changes in the strength of the earth's field. The belt's outer edge is probably controlled by hydromagmagnetic waves which change the protons' magnetic moment and scatter them into the atmosphere. As far as we know now, these protons are made from neutrons and from no other sources.

Medium-energy protons (5 Mev $< E_p <$ 30 Mev) seem to be made from neutron decay by two different processes: In the inner part of the inner belt, for $L < 1.6$ where the spectrum is rather flat, the protons are quite certainly due to neutrons made by galactic cosmic-ray protons. Out farther, for $L > 1.7$, the steeper spectrum and larger fluxes are quite certainly due to the decay of neutrons made by solar protons striking the polar regions at the times of some solar flares.

The newly discovered low-energy protons (0.5 kev $< E_p <$ 1 Mev) might either be made by some quite new and different and not understood source, or they might be accelerated from a lower-energy source (such as slow neutron-decay protons or knock-on protons) to the observed energies. One thing is quite certain: neutron-decay protons do not supply a large fraction of the total energy in these protons directly.

The origin of the trapped electrons in the inner belt is not as well established because there have been few experiments to study their characteristics. The flux is not known to an order of magnitude and the spectrum is also uncertain, but it would appear that the energy spectrum goes up past 1 Mev and the flux is of the order 10^7 electrons/cm^2-sec. From neutrons we would expect a flux of about 10^7 and a spectrum ending at 780 kev. The high-energy electrons ($E > 780$ kev) are not produced by neutron decay directly, but it would seem that some acceleration processes operating on neutron-decay electrons to increase their energy is the most probable source mechanism for these electrons. If we use an accelerating process operating on low-energy electrons to generate the observed electrons, the requirements on the accelerating process are considerably increased.

Our conclusion is that the inner-belt protons of $E > 5$ Mev are produced by neutron decay from galactic and solar protons, but the $E < 1$ Mev protons are of different and unknown origin. The inner-belt electrons are probably made by a combination of neutron decay and some uncertain acceleration process.

REFERENCES

1. Van Allen, J. A., and Ludwig, G. H., et al., "Observation of High Intensity Radiation by Satellites 1958 Alpha and Gamma," *Jet Propulsion*, **28**(9): 588–592, September 1958.
2. ———, "First Public Lecture on the Discovery of the Geomagnetically-Trapped Radiation," State Univ. of Iowa, SUI 60-13, 1960.
3. Störmer, C., "The Polar Aurora," Oxford: Clarendon Press, 1955.
4. Van Allen, J. A., "Dynamics, Composition and Origin of the Geomagnetically-Trapped Corpuscular Radiation," State Univ. of Iowa, SUI 61-19, August 1961.
5. Chapman, S., and Bartels, J., "Geomagnetism," Oxford: Clarendon Press, 1940.
6. Post, R. F., "Controlled Fusion Research—An Application of the Physics of High Temperature Plasmas," *Rev. Mod. Phys.*, **28**(3): 338–362, July 1956.
7. Christofilos, N. C., "The Argus Experiment," *J. Geophys. Res.*, **64**(8): 869–875, August 1959.
8. Vernov, S. N., and Chudakov, A. E., "Terrestrial Corpuscular Radiation and Cosmic Rays," in: *Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January* 1960, Amsterdam: North-Holland Publ. Co., 1960, pp. 751–796.
9. Krassovsky, V. I., and Shklovsky, I. S., et al., "On Fast Corpuscles of the Upper Atmosphere," in: *Proc. Internat. Conf. of Cosmic Radiation, Moscow, July* 1959, Moscow, Academie Nauk, 1960, Vol. 3, pp. 59–63.
10. Van Allen, J. A., McIlwain, C. E., and Ludwig, G. H., "Radiation Observations with Satellite 1958:," *J. Geophys. Res.*, **64**(3): 271–286, March 1959.
11. ———, ———, and ———, "Satellite Observations of Electrons Artificially Injected into the Geomagnetic Field," *J. Geophys. Res.*, **64**(8): 877–891, August 1959.

12. ———, and FRANK, L. A., "Radiation around the Earth to a Radial Distance of 107,400 km," *Nature*, **183**(4659): 430–434, February 14, 1959; Also State Univ. of Uowa, SUI 59–2.
13. ———, and ———, "Radiation Measurements to 658,300 km with Pioneer IV," *Nature*, **184**(4682): 219–224, July 25, 1959; Also State Univ. of Iowa, SUI 59–18, August 1959.
14. ALFVÉN, H., "Cosmical Electrodynamics," Oxford, Clarendon Press, 1950.
15. SPITZER, J., Jr., "Physics of Fully Ionized Gases," New York: Interscience Publ., 1956.
16. CHRISTOFILOS, N. C., "Trapping and Lifetime of Charged Particles in the Geomagnetic Field," Univ. Calif. Radiation Lab. Rept. UCRL–5407, November 28, 1958.
17. GOLD, T., "Plasma and Magnetic Fields in the Solar System," *J. Geophys. Res.*, **64**(11): 1665–1674, November 1959.
18. HERLOFSON, N., "Diffusion of Particles in the Earth's Radiation Belts," *Phys. Rev. Letters*, **5**(9): 414–416, November 1, 1960.
19. ROSENBLUTH, M. N., and LONGMIRE, C. L., "Stability of Plasmas Confined by Magnetic Fields," *Ann. of Physics*, **1**(2): 120–140, May 1957.
20. NORTHROP, T. G., and TELLER, E., "Stability of the Adiabatic Motion of Charged Particles in the Earth's Field," *Phys. Rev.*, **117**(1): 215–225, January 1, 1960.
21. GALL, R., and LIFSHITZ, J., "Temporary Capture of Cosmic Ray Particles and their Contribution to the High Intensity Belts," *Nuovo Cimento*, **15**(2): 233–245, January 16, 1960.
22. BRIDGE, H. S., and DILWORTH, C., et al., "Direct Observations of the Interplanetary Plasma," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961. II. *Main Sessions*, Tokyo: Physical Society of Japan, 1962, pp. 553–560.
23. CHANG, C. C., "Outer Van Allen Belts and Neutral Points on Interface Between Solar Wind and Geomagnetic Field," *Nature*, **194**(4827): 424–426, May 5, 1962.
24. AXFORD, W. I., and HINES, C. O., "A Unifying Theory of High-Latitude Geophysical Phenomena and Geomagnetic Storms," *Can. J. Phys.*, **39**(10): 1433–1464, October 1961.
25. HESS, W. N., and STARNES, A. J., "Measurement of the Neutron Flux in Space," *Phys. Rev. Letters*, **5**(2): 48–50, July 15, 1960.
26. ———, CANFIELD, E. H., and LINGENFELTER, R. E., "Cosmic-Ray Neutron Demography," *J. Geophys. Res.*, **66**(3): 665–677, March 1961.
27. ———, and PATTERSON, H. W., et al., "Cosmic-Ray Neutron Energy Spectrum," *Phys. Rev. Letters*, **116**(2): 445–457, October 15, 1959.
28. STUART, G. W., "Satellite-Measured Radiation," *Phys. Rev. Letters* **2**(10): 417–418, May 15, 1959.
29. FITE W. L., STEBBINGS, R. F., et al., "Ionization and Charge Transfer in Proton-Hydrogen Atom Collisions," *Phys. Rev.* **119**(2): 663-668, July 15, 1960.
30. OREAR, J., ROSENFELD, A. H., and SCHLUTER, R. A. (Comp.), "Nuclear Physics; A Course Given by Enrico Fermi at the University of Chicago," Rev. Ed., Chicago: University of Chicago Press, 1950.
31. FREDEN, S. C., and WHITE, R. S., "Protons in the Earth's Magnetic Field," *Phys. Rev. Letters*, **3**(1): 9–11, July 1, 1959.
32. ———, and ———, "Trapped Proton and Cosmic Ray Albedo Neutron Fluxes," *J. Geophys. Res.*, **67**(1): 25–29, January 1962.
33. ———, and ———, "Particle Fluxes in the Inner Radiation Belt," *J. Geophys. Res.*, **65**(5): 1377–1383, May 1960.
34. ARMSTRONG, A. H., and HARRISON, F. B., et al., "Charged Particles in the Inner Van Allen Radiation Belt," *J. Geophys. Res.*, **66**(2): 351–357, February 1961.
35. NAUGLE, J. E., and KNIFFEN, D. A., "Flux and Energy Spectra of the Protons in the Inner Van Allen Belt," *Phys. Rev. Letters*, **7**(1): 3–6, July 1961.
36. HECKMAN, H. H., and ARMSTRONG, A. H., "Energy Spectrum of Geomagnetically Trapped Protons," *J. Geophys. Res.*, **67**(4): 1255–1262, April 1962.
37. SINGER, S. F., "Trapped Albedo Theory of the Radiation Belt," *Phys. Rev. Letters*, **1**(5): 181–183, September 1, 1958.
38. ———, "Latitude and Altitude Distribution of Geomagnetically Trapped Protons," *Phys. Rev. Letters*, **5**(7): 300–303, October 1, 1960.
39. LENCHEK, A. M., and SINGER, S. F., "Geomagnetically Trapped Protons from Cosmic-Ray Albedo Neutrons," *J. Geophys. Res.*, **67**(4): 1263–1288, April 1962.
40. HESS, W. N., "Van Allen Belt Protons from Cosmic-Ray Neutron Leakage," *Phys. Rev. Letters*, **3**(1): 11–13, July 1, 1959.
41. HARRIS, I., and PRIESTER, W., "Theoretical Models for the Solar-Cycle Variation of the Upper Atmosphere," NASA Technical Note D–1444, August 1962.
42. RAY, E. C., "On the Theory of Protons Trapped in the Earth's Magnetic Field," *J. Geophys. Res.*, **65**(4): 1125–1134, April 1960.
43. YOSHIDA, S., LUDWIG, G. H., and VAN ALLEN, J. A., "Distribution of Trapped Radiation," *J. Geophys. Res.*, **65**(3): 807–813, March 1960.
44. FAN, C. Y., MEYER, P., and SIMPSON, J. A., "Dynamics and Structure of the Outer Radiation Belt," *J. Geophys. Res.*, **66**(9): 2607–2640, September 1961.
45. WELCH, J. A., and WHITAKER, W. A., "Theory of Geomagnetically Trapped Electrons from an Artificial Source," *J. Geophys. Res.*, **64**(8): 909–922, August 1959.
46. DRAGT, A. J., "Effect of Hydromagnetic Waves on the Lifetime of Van Allen Radiation Protons," *J. Geophys. Res.*, **66**(6): 1641–1649, June 1961.
47. WENTZEL, D. G., "Hydromagnetic Waves and the Trapped Radiation. Part 2. Displacements of the Mirror Points," *J. Geophys. Res.*, **66**(2): 363–369, February 1961.

48. PARKER, E. N., "Effect of Hydromagnetic Waves in a Dipole Field on the Longitudinal Invariant," *J. Geophys. Res.*, **66**(3): 693–708, March 1961.
49. HAMLIN, D. A., and KARPLUS, R., et al., "Mirror and Azimuthal Drift Frequencies for Geometrically Trapped Particles," *J. Geophys. Res.*, **66**(1): 1–4, January 1961.
50. FERMI, E., "On the Origin of Cosmic Radiation," *Phys. Rev.*, **75**(8): 1169–1174, April 15, 1949.
51. MCILWAIN, C. E., "Coordinates for Mapping the Distribution of Magnetically Trapped Particles," *J. Geophys. Res.*, **66**(11): 3681–3691, November 1961.
52. ROTHWELL, P., and MCILWAIN, C. E., "Magnetic Storms and the Van Allen Radiation Belts—Observations from Satellite 1958: (Explorer IV)," *J. Geophys. Res.*, **65**(3): 799–806, March 1960.
53. ARMSTRONG, A. H., and HECKMAN, H. H., "Flux and Spectrum of Charged Particles in the Lower Van Allen Belt," *Bull. Amer. Phys. Soc.*, **6**(4): 361, June 22, 1961 (Abstract).
54. PIZZELLA, G., MCILWAIN, C. E., and VAN ALLEN, J. A., "Time Variations of Intensity in the Earths Inner Radiation Zone, October 1959 through December 1960," *J. Geophys. Res.*, **67**(4): 1235–1254, April 1962.
55. BAME, S. J., and CONNER, J. P., et al., "Protons in the Outer Van Allen Belt," *J. Geophys. Res.*, **68**(1): 55–63, January 1963.
56. LENCHEK, A. M., "On the Anomalous Component of Low-Energy Geomagnetically Trapped Protons," *J. Geophys. Res.*, **67**(6): 2145–2157, June 1962.
57. WEBBER, W. R., "Time Variations of Low Energy Cosmic Rays during the Recent Solar Cycle," in: *Progress in Elementary Particle and Cosmic Ray Physics*, ed. by J. G. Wilson and S. A. Wouthuysen, Amsterdam: North-Holland Publ. Co., Vol. 6, 1962 (in press).
58. FREEMAN, J. W., "Detection of an Intense Flux of Low-Energy Protons or Ions Trapped in the Inner Radiation Zone," *J. Geophys. Res.*, **67**(3): 921–928, March 1962.
59. HOLLY, F. E., ALLEN, L., and JOHNSON, R. G., "Radiation Measurements to 1500 Kilometers Altitude at Equatorial Latitudes," *J. Geophys. Res.*, **66**(6): 1627–1639, June 1961.
60. MANN, L. G., BLOOM, S. D., and WEST, H. I., Jr., "The Electron Spectrum from 80 to 1258 Kev Observed on Discoverer Satellites 29 and 31," Paper presented at the 3rd Internat. Space Sci. Sympos., Washington, May 1962.
61. VERNOV, S. N., and SAVENKO, I. A., et al., "Discovery of the Internal Radiation Belt at a Height of 320 km in the Region of the South Atlantic Magnetic Anomaly," *Doklady Akademii Nauk SSSR*, **140**(5): 1041–1044, October 11, 1961. Translation in *Soviet Phys.—Doklady*, **6**(10): 893–896, April 1962.
62. KELLOGG, P. J., "Electrons of the Van Allen Radiation," *J. Geophys. Res.*, **65**(9): 2705–2713, September 1960.
63. LENCHEK, A. M., SINGER, S. F., and WENTWORTH, R. C., "Geomagnetically Trapped Electrons from Cosmic Ray Albedo Neutrons," *J. Geophys. Res.*, **66**(12): 4027–4046, December 1961.
64. HESS, W. N., and POIRIER, J. A., "Energy Spectrum of Electrons in the Outer Radiation Belt," *J. Geophys. Res.*, **67**(5): 1699–1709, May 1962.
65. WENTWORTH, R. C., MACDONALD, W. W., and SINGER, S. F., "Lifetimes of Trapped Radiation Belt Particles Determined by Coulomb Scattering," *Physics of Fluids*, **2**(5): 499–509, September–October 1959.
66. O'BRIEN, B. J., and LANGHLIN, C. D., et al., "Measurements of the Intensity and Spectrum of Electrons at 1000-Kilometer Altitude and High Latitudes," *J. Geophys. Res.*, **67**(4): 1209–1226, April 1962.
67. CLADIS, J. B., and CHASE, L. F., et al., "Energy Spectrum and Angular Distributions of Electrons Trapped in the Geomagnetic Field," *J. Geophys. Res.*, **66**(8): 2297–2312, August 1961.
68. O'BRIEN, B. J., and VAN ALLEN, J. A., et al., "Absolute Electron Intensities in the Heart of the Earth's Outer Radiation Zone," *J. Geophys. Res.*, **67**(1): 397–403, January 1962.
69. ROSSER, W. G. V., and O'BRIEN, B. J., et al., "Electrons in the Earth's Outer Radiation Zone," Paper presented at 43rd Annual Meeting Amer. Geophys. Union, Washington, April 1962.
70. ALLER, L. H., "Astrophysics; the Atmospheres of the Sun and Stars," New York: Ronald Press, 1953.
71. BISWAS, S., FICHTEL, C. E., and GUSS, D. E., "A Study of the Hydrogen, Helium and Heavy Nuclei in the November 12, 1960 Solar Cosmic Ray Event," *Phys. Rev.*, **128**(6): 2756–2771, December 1962.

LIFETIME AND TIME HISTORIES OF TRAPPED RADIATION BELT PARTICLES

WILMOT N. HESS

Goddard Space Flight Center

In this paper recent advances in understanding the lifetimes of trapped radiation belt particles are surveyed. The problem of particle lifetimes has been one of the most important and perplexing in radiation belt work for several years. As a result of the past year's work, a good deal more is known about the natural Van Allen radiation belt and also the characteristics of the artificial belts produced by high altitude nuclear explosions. This increased knowledge is based on both observations and calculations of the artificial belts and Van Allen belt. This better understanding of particle lifetimes is a substantial step forward in forming an overall picture of the radiation belt in that we now have a basis for selecting source and loss processes more realistically and therefore making better models of the belt.

Electron lifetimes based on data collected from the high altitude explosions of the U.S. and U.S.S.R., and proton lifetimes in the Van Allen belt are also discussed.

INTRODUCTION

The first of the four radiation belts made in 1962 resulted from the explosion of a 1.4 megaton nuclear bomb by the U.S. 400 km above Johnston Island in the Pacific on July 9. This generated a new belt of trapped electrons resulting mainly from the decay of fission fragments, but possibly involving other processes as well. Measurements by Injun, 1961o2,[1] which had been in orbit about a year, showed clearly that at 1000 km a marked increase in electron flux occurred. A GM counter on Injun shielded with about 4 gm/cm^2 of Pb, counting electrons of several Mev via bremsstrahlung, mapped out the high energy electron distribution in space up to 1000 km. A GM counter on Traac, 1961$\alpha\eta$2,[2] shielded by 0.66 gm/cm^2 and counting electrons of $E>1.6$ Mev also observed the early effects and mapped the particle distribution. A GM counter on Ariel, 1962o1,[3] showed that the explosion produced high energy electrons out to $L\sim5$. The most complete mapping of the belt was made by Telstar, 1962$\alpha\epsilon$1,[4] which went considerably higher than the others to an apogee of 5630 km. Telstar unfortunately did not get launched until the day after the explosion. Telstar had a four channel solid state detector counting electrons out past $L=2$.

Comparing the several detector count rates at about 1000 km showed that the new electrons had roughly an equilibrium fission energy spectrum identifying them as coming from β-decay of fission fragments.[5]

The difference between the more extensive Telstar flux map and the Injun flux map has caused some questions in the past, but it is starting to be understood now in terms of variations of the electron energy with L. Both the Injun and Telstar contours are correct, but they must be interpreted in terms of different energy electrons.

Most of the information about the electron lifetimes obtained recently has come from watching the decay of this population of particles from Starfish.

$L>1.7$

For $L<1.7$ the decay of the electrons introduced by the Starfish explosion is quite slow and appears to be controlled by the atmosphere. Coulomb scattering of the electrons by the atmospheric atoms will result in changing the direction of

motion of the electron and therefore changing the pitch angle, α(the angle between **B** and **v**, the electron velocity). This scattering will result in changing the mirror point altitude. This can be seen from inspection of the mirror equation

$$\frac{\sin^2 \alpha}{B} = \frac{1}{B_m} .$$

If at some point with magnetic field B, the pitch angle α is changed, then B_m the value of the magnetic field at the mirror point is also changed. This means the mirror altitude is changed. A series of Coulomb scatters will move a particle mirror point up and down a field line, but out of this process a net loss of particles into the atmosphere will occur. This loss can be understood

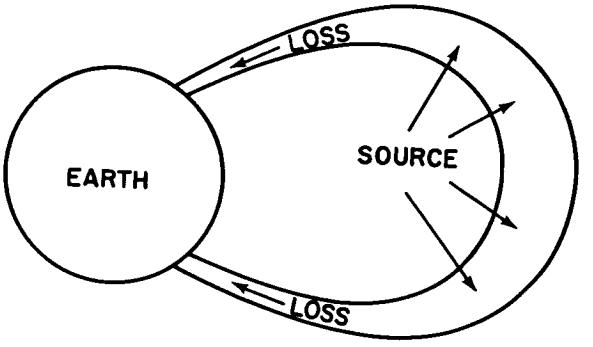


FIGURE 1.—The steady state model of the radiation belt.

physically at low altitudes. If a scatter occurs vary near a particle's mirror point, it can only lower the mirror point. At, and only at, the mirror point the particle's motion is perpendicular to the field line so any scattering at this point either up or down which makes the motion not perpendicular to the filed line can only lower the mirror point (figure 1).

The effect of repeated Coulomb collisions can be calculated by using a Fokker–Plank equation[6] which describes how a distribution of particles $U(B, t, E)$ in a tube of force changes with time as the result of Coulomb collisions:

$$\frac{\partial U}{\partial t} = -\frac{\partial}{\partial B}\left[\langle \delta B \rangle U\right] - \frac{\partial}{\partial E}\left[\langle \delta E \rangle U\right]$$

$$+ \frac{1}{2}\frac{\partial^2}{\partial B^2}\left[\langle (\delta B)^2 \rangle U\right] .$$

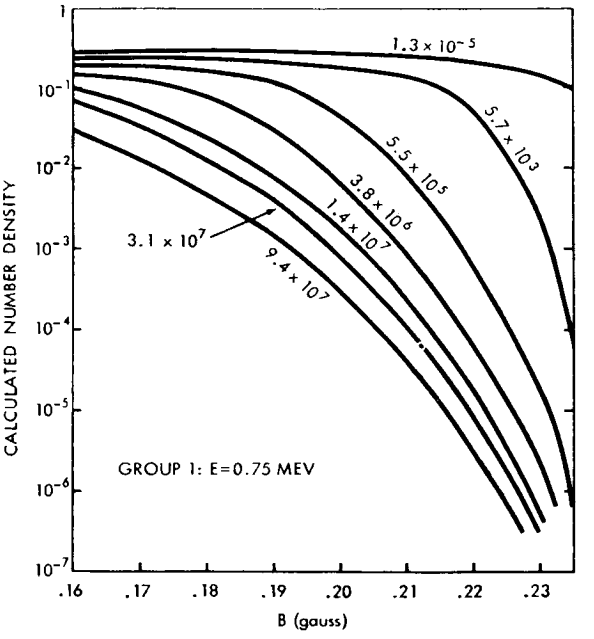


FIGURE 2.—Calculated number density of $E=0.75$ Mev electrons for $L=1.25$ for various times after injection (reference 2).

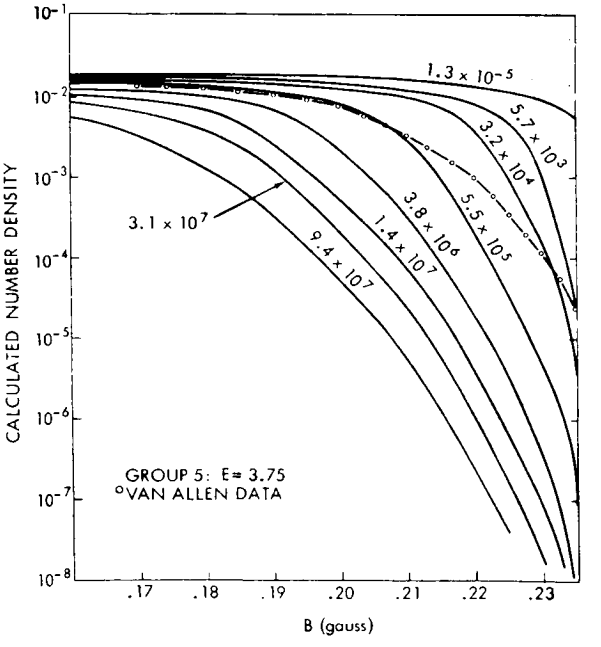


FIGURE 3.—Calculated number density of $E=3.75$ Mev electrons for $L=1.25$ for various times after injection (reference 7). The dotted curve labeled Van Allen Data is a smooth curve drawn through the Injun data (reference 1) and is for $t \sim 10^6$ sec.

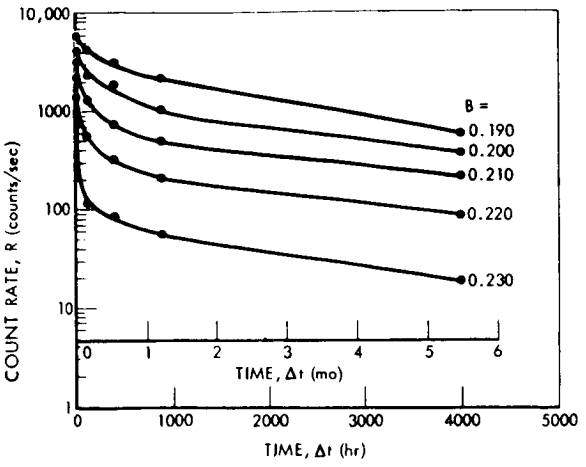


FIGURE 4.—Experimental decay curves of the Starfish electrons for $1.25 \leq L \leq 1.35$ from Injun I and Injun III (reference 9).

Where the terms in $\langle\ \rangle$ are the average time rates of time of the quantities. This has been integrated on a computer.[7] A tube of force is divided into 100 space cells of equal ΔB and the fission energy spectrum into 9 groups and an intial mirror point distribution is selected. The $\langle\ \rangle$ terms are evaluated for each cell and group by averaging the scattering events over a bounce period. In doing this the atmospheric model taken to calculate the scattering was that of Harris and Priester[8] averaged in local time and in longitude. The variation of the particles' altitude with longitude is quite important in doing this average. Then the equation is integrated a time step, dt, first for scattering and then for energy loss, and the process repeated a large number of times in order to see the distribution change.

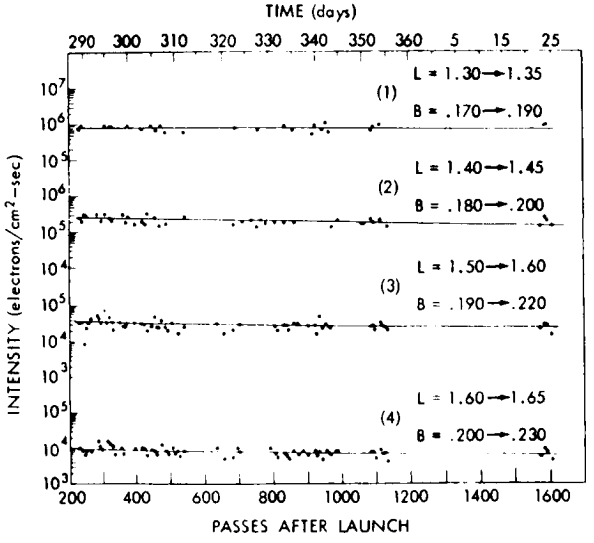


FIGURE 5.—Experimental decay curves of the Starfish electrons for Alouette (reference 10).

The time history of the population for $L=1.25$ is shown in figure 2 for $E=0.75$ Mev and in figure 3 for $E=3.75$ Mev. As is expected, the first particles to be lost are the ones at high B (or low altitude). Gradually the decay rate slows and the spatial distribution eventually reaches an equilibrium shape. For the equilibrium situation, scattering down the line is nearly balanced by scattering up the line so the decay proceeds slowly, being dominated by the scattering rate at the equator. Van Allen, Frank and O'Brien[9] on Injun I and Injun III (1962$\beta\tau$2) have measured the decay of the Starfish electrons over a period of 4000 hrs

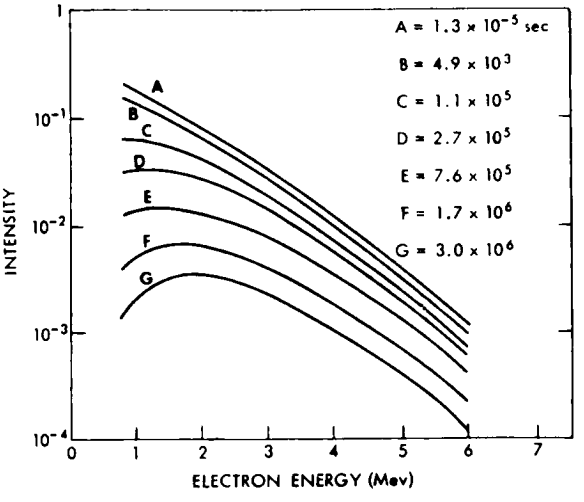


FIGURE 6.—The calculated electron energy spectrum for $L=1.18$ and $B=0.20$ at various times after injection of a fission spectrum (reference 7).

(figure 4). Also, McDiarmid, Burrows, Budzinski and Rose[10] have measured the decay of the Starfish electrons of $E>3.9$ Mev by a GM counter on the Alouette satellite and this is shown in figure 5. They find a decay constant of about 550 days for $B=0.18$ and $L=1.32$. The characteristics of the experimentally observed decay agreed with that which is expected from atmospheric decay.

During the process of atmospheric scattering

the energy spectrum of the electrons changes. The lower energy electrons are more easily scattered and therefore lost first. Because of this, the fission energy spectrum hardens with time until an equilibrium spectrum is developed which has a peak at about 1.5 Mev. This process of spectral change is shown in figure 6 for $L=1.18$, as calculated by Welch, Kaufmann and Hess.[7]

$L>1.7$

The time history of the electrons for large L values after Starfish is quite different than for $L<1.7$. The solid state detector on Telstar counting electrons of $E\sim0.5$ Mev showed the time history given in figure 7[11] starting on July 10, which was the Telstar launch day and the day after Starfish. These decay curves are each for a narrow range of B of about 0.60 gauss. Figure 7 shows very clearly the time decay of a transient particle population down to something resembling a steady state population in a period of about three months. So even though Telstar did not observe the particle populations before the Starfish event, one can tell from its record that a large transient population was produced out past $L=2.5$ at about the time of Starfish.

Brown and Gabbe[4] calculated decay constants for $L>2.0$ based on the initial slopes of the figure 7 decay curves. Figure 8 shows values of the decay constant τ at different L values. At $L\sim2.5$ the electron's mean life is only a few days. This is

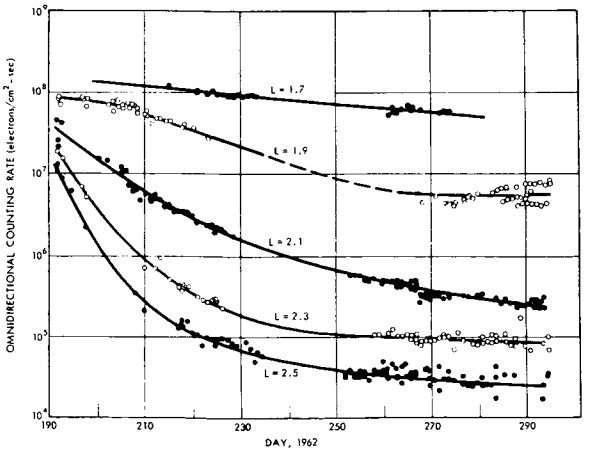


FIGURE 7.—Experimental time histories of electrons of $E\sim0.5$ Mev as measured by Telstar (reference 11) for $L>1.7$ starting on July 10, 1962.

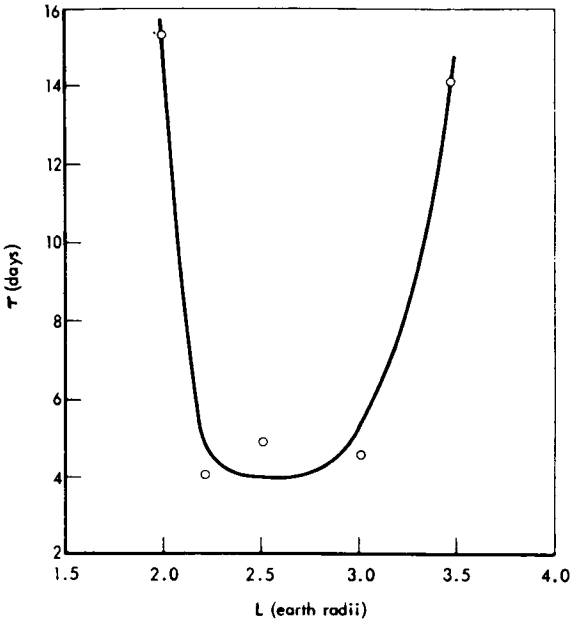


FIGURE 8.—Experimental decay constants for Starfish electrons as measured in July 1962 by Telstar (reference 4).

very different from the particle lifetime of years at $L\sim1.4$. For $L>1.7$ the decay rate gets markedly shorter than values expected from atmospheric decay. In figure 8 the increase in τ for $L>3$ may not be physically real, but may result from the fact that not many new particles were injected at these large L values by Starfish. In a region where no particles were added and a steady state population existed there would be no decay observed and $\tau\rightarrow\infty$.

There is no good explanation as to why the electron lifetimes are so short for $L>1.7$. The process responsible for this seems to have a quite sudden onset at $L\sim1.7$, and by $L\sim2.2$, the electron lifetimes has been decreased roughly three orders of magnitude from that expected from atmospheric decay. The best candidate for this loss process is magnetic scattering. Welch and Whitaker[12] suggested that magnetic disturbances might break down the first adiabatic invariant and the particles would be magnetically scattered and lost. This requires frequencies of the disturbances of nearly a megacycle. Parker[13] has suggested that breakdown of the longitudinal invariant due to magnetic disturbances moving the

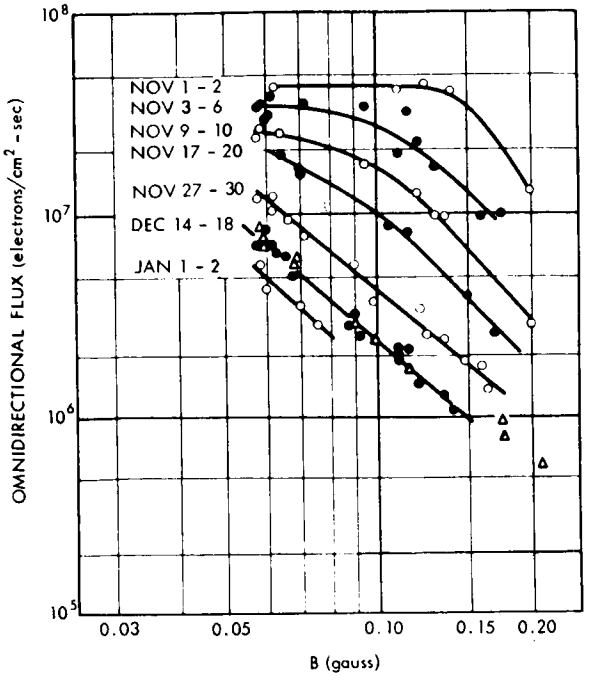


FIGURE 9.—Experimental time histories of electrons of $E>1.9$ Mev from Explorer XV (reference 11) for $L=1.75$. The triangles are for October 28 through 31 before the third U.S.S.R. explosion.

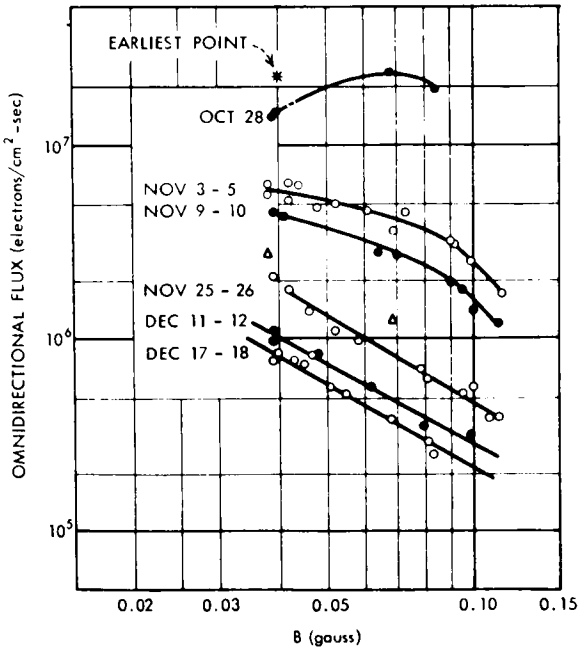


FIGURE 10.—Experimental time histories of electrons of $E>1.9$ Mev from Explorer XV (reference 4) for $L=2.0$.

particles mirror point would remove particles from the belt. This effect can be caused by waves of a few cps and therefore might be easier to make work. But whether such a process would have the sharp variation with L observed is uncertain. There is now no quantitative explanation for the short lifetime.

Brown and Gabbe[11] also observed the particles

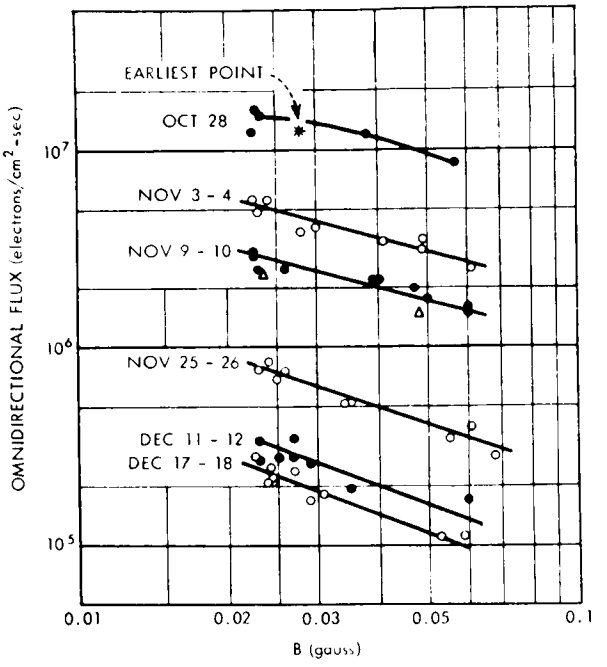


FIGURE 11.—Experimental time histories of electrons of $E>1.9$ Mev from Explorer XV (reference 4) for $L=2.4$.

injected into the belt by the three USSR high altitude nuclear explosions of October 22, October 23, and November 1, 1962. Their instruments on Telstar studied the three explosions. Also, their instruments on the Explorer XV (1962$\beta\lambda$2) satellite launched on October 27 observed the October 28 and November 1 USSR explosions. The time histories for this period for their detector of $E>1.9$ Mev on Explorer XV are shown in figures 9, 10, and 11 for $L=1.75$, 2.0 and 2.4 respectively. These time histories show a quite similar decay to the Telstar decay curves after Starfish. There is an initial redistribution of the flux along a field line for the low L lines followed by a decay with rather similar τ values to figure 8. It seems that this decay is due to a usual condition in the mag-

netosphere and does not depend upon solar storms or other occasional events although such events may also be important.

PROTONS

It is well established that the high energy protons in the inner belt have long lifetimes. Freden and White[14] in an analysis of proton tracks in emulsion stacks deduced the proton lifetimes shown in figure 12. But in spite of these long lifetimes, time changes have been seen in the heart of the inner belt. At $L=1.3$ a gradual increase of a factor of three in the protons flux was observed during the year 1960.[15] This flux change can be understood in terms of the change in exospheric density during the solar cycle.[16] We have calculated the solar cycle changes expected for $L=1.25$ for several values of B.[26] We start with the continuity equation:[17]

$$\frac{dN(E)}{dt} = S(E) - L(E) + \frac{\partial}{\partial E}\left[N(E)\frac{dE}{dx}\frac{dx}{dt}\right] ,$$

but we will not assume equilibrium, so

$$\frac{dN(E)}{dt} \neq 0 .$$

For the source term $S(E)$ we use the energy data from Hess, Canfield and Lingenfelter[18] and the time behavior from McDonald and Webber:[19]

$$S(E) = \frac{.8E^{-2.0}}{\gamma^{v\tau}} f(t) .$$

For the loss term, following Freden and White,[14] we take

$$L = N(E)\sum_i n_i \sigma_i ,$$

where n_i is the atomic density of component i and σ_i is the inelastic cross section of component i.

We take the rate of energy loss,

$$\frac{dE}{dx} = \left(\frac{\rho}{\rho_N}\right)\left.\frac{dE}{dx}\right|_N$$

where $(dE/dx)_N$ is the energy loss rate for NTP air of density ρ_N given in Aron, Hoffman, and Williams[20] and ρ is the actual average atmospheric density encountered by the protons. We have used the Harris and Priester[8] atmosphere and have averaged in local time and in longitude for each value of B and L and averaged along the bounce of the particle. These average densities ρ vary with time of the solar cycle.

Using this $\rho(t)$ we have calculated $N(E, t)$ from the equation for $dN(E)/dt$. Plots of $N(E, t)$ are shown in figure 13 for several energies starting from $N=0$ at $t=0$. For high altitude >2000 km the time variations are small because the protons

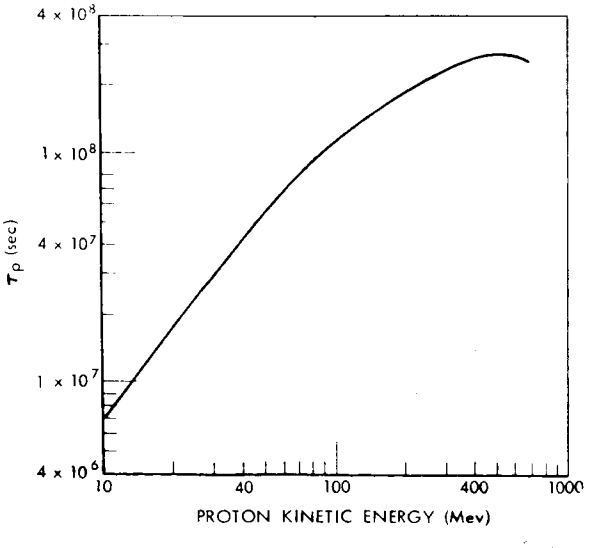


FIGURE 12.—Decay constants of high energy Van Allen belt protons calculated by Freden and White (reference 14).

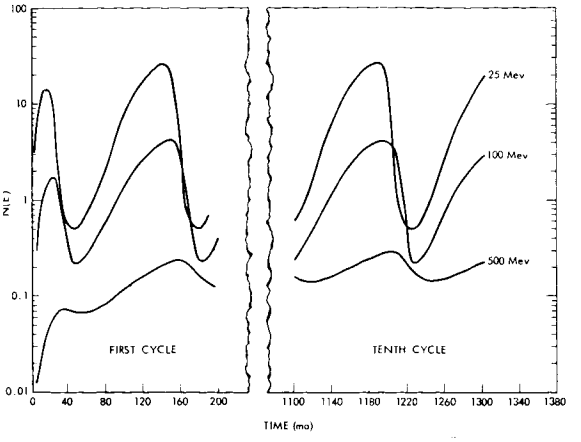


FIGURE 13.—Time histories of protons of different energies for $L=1.25$, $B=0.209$, $h^{min}=400$ km starting from no flux and following the flux buildup through the 1st solar cycle and also the tenth solar cycle (Blanchard and Hess, 1963, reference 26).

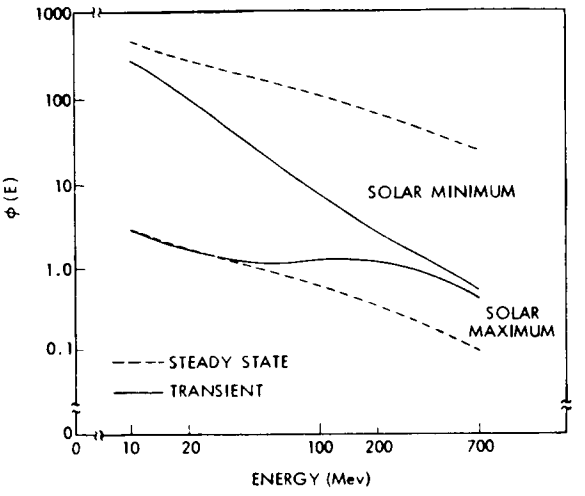


FIGURE 14.—The proton energy spectrum at solar max and solar min for $L=1.25$, $B=0.199$, $h^{min}=500$ km. The dashed curves are the spectra that would exist if the solar max and solar min atmospheres existed for long enough times for the problem to come to equilibrium (Blanchard and Hess, 1963, reference 26).

lifetime is long compared to the solar cycle. For low altitudes >200 km the atmosphere does not change much with the solar cycle, so no large flux variations are seen. But for middle altitudes important changes do take place. The energy spectrum also changes significantly during the solar cycle. Figure 14 shows the spectrum at solar maximum and solar minimum for $B=0.199$. Figure 15 shows the amplitude of the solar cycle flux variation for $L=1.25$ for values of B and E. These fluctuations must take place if the protons have long lives because we know the exospheric density changes with time. The details of the changes will depend on the detailed characteristics of the atmosphere. The Harris and Priester atmosphere gives the best current values of the effect. There is not currently available very complete experimental information to use to compare with these calculations. When such data becomes availble it may help improve the understanding of the upper atmosphere.

Filz, Holeman, and Yagoda[21] have analyzed nuclear emulsions flown on several low altitude satellites and found that the proton flux for $E_p>55$ Mev was increased about a factor of 5 at 400 km shortly after Starfish. This is probably due to a redistribution of a small fraction of the trapped high energy protons by some process related to the explosion. This increase in proton population was temporary and the flux decayed back to the steady state value in a few months. Filz, Holeman, and Yagoda give the time for the 55 Mev protons to decay by a factor of 10 at 420 km as 110 days or a mean life $\tau\sim5\times10^6$ sec.

If we assume most of the flux was encountered in the South Atlantic magnetic anomaly at $B=0.210$, $L\sim1.3$ and using the Harris and Priester $S=100$ averaged atmospheric density $\rho=10^6$, we get the lifetime for a 70 Mev proton from Freden and White:[14]

$$\tau = \frac{7.5\times10^{10}}{\rho} E^{1.31} = 2\times10^7 \text{ sec} .$$

This gives only fair agreement with the experimental proton lifetime. It is also uncertain whether the decay at different altitudes proceeded at different rates, as it should according to the theory.

It is well established that cosmic ray neutrons produce the high energy protons in the inner Van Allen zone. Measurements show that these high energy protons are not present in the outer zone although there the neutron source will still make them. Some additional loss process must operate to remove them more rapidly from the outer zone than from the inner zone. Garmire[22] has recently analyzed the count rate of the Cerenkov counter on the Explorer XI (1961ν1) satellite to study

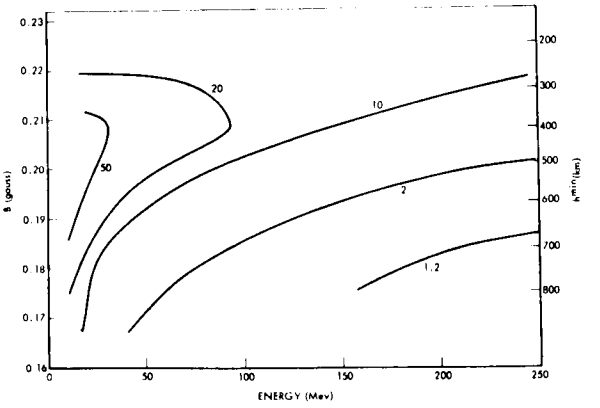


FIGURE 15.—The amplitude of the solar min/solar max flux ratio for protons of different energies at different values of B for $L=1.25$ (Blanchard and Hess, 1963, reference 26).

protons of $E_p > 350$ Mev. This is part of the detector used to study high energy γ-rays in space, but it also responds to trapped high energy protons. Garmire compared his count-rate contours with those of McIlwain[23] for $E_p > 43$ Mev to show that there are relatively fewer high energy protons at large L values. One process which could produce such a situation would be a breakdown of the first adiabatic invariant $J\mu$ involving the constancy of the particles magnetic moment. Welch and Whitaker[12] first suggested that hydromagnetic waves might break down this invariant. Lenchek and Singer[24] following calculations of Garren, et al.[25] proposed that for high enough energy protons in a static dipole field the Alfven approximation might not hold and the particles would be lost.

Garmire[22] has numerically computed trapped-proton orbits in a dipole field and also in a Finch and Leaton field and has obtained an effective cutoff energy for trapped protons in both fields as a function of position in the field (figure 16). It happens, interestingly, that the cutoff energies are considerably lower for the Finch and Leaton field. Garmire's comparison of the cutoff energy calculations with the experimental data shows that this process of the breakdown of $J\mu$ very likely contributes to the loss of high energy protons at large L. This might even be the sole loss process if the variation with L is rapid enough. If this process is responsible for the loss of high energy outer zone protons then the outer edge of the inner zone should not change appreciably with time in the solar cycle because this loss process is not time dependent.

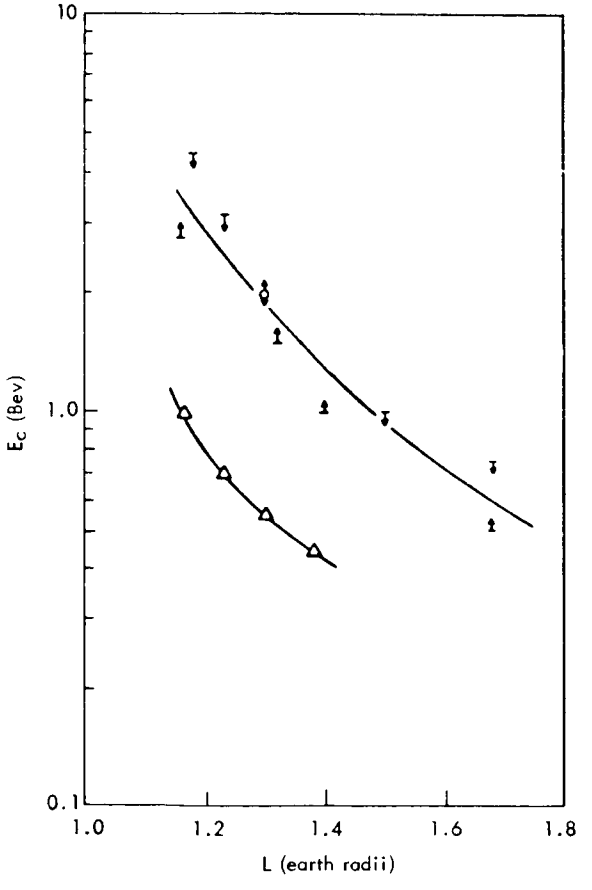


FIGURE 16.—Computed values of the cutoff energy Ec for high energy trapped protons for a dipole field and for a Finch and Leaton field (reference 22).

CONCLUDING REMARKS

Last year, for the first time, we have obtained definitive information on trapped electron lifetimes:

1. For $L < 1.7$ the dominant loss process is atmospheric scattering and the particle lifetimes are long. At $L = 1.3$ the electron's mean life is about a year. The scattering process creates an equilibrium flux distribution along a field line and an equilibrium energy spectrum;

2. For $L > 1.7$ the electrons' lifetimes get much shorter than that expected from atmospheric scattering. At $L = 2.5$ the electrons mean life is about one week. The loss appears to be gradual in time and an equilibrium flux distribution appears for $1.7 < L < 2.5$. The nature of the loss process operating here is not understood.

New information also helps understanding of proton time histories. Inner zone protons made by neutron decay have long lifetimes, but they will show time changes too:

3. At several hundred kilometers altitude calculations show that the proton population should change more than an order of magnitude during the solar cycle due mostly to changes in the exospheric density. This process also changes the proton energy spectrum. At high altitudes the proton lifetime is long compared to the solar cycle and the effect is averaged out. Fragmentary experimental information tends to confirm this.

4. A direct measurement at 420 km of the lifetime of $E>55$ Mev protons displayed by the Starfish explosion agrees only fairly well with the lifetime value calculated on the basis of atmospheric slowing down.

5. High energy protons do not appear in the outer zone. Calculations show now that they may be lost because their Störmer orbits are not trapped. The non-dipolar nature of the field is important in carrying out these calculations. The result of the calculations agrees reasonably well with experimental data from Explorer XI. This loss process would be constant in time and so the outer edge of the inner belt should not vary much during the solar cycle.

REFERENCES

1. O'Brien, B. J., Laughlin, C. D., and Van Allen, J. A., "Geomagnetically Trapped Radiation Produced by a High-Altitude Nuclear Explosion on July 9, 1962," *Nature*, **195**(4845): 939–942, September 8, 1962.
2. Pieper, G. F., Williams, D. J., and Frank, L. A., "Traac Observations of the Artificial Radiation Belt from the July 9, 1962, Nuclear Detonation," *J. Geophys.*, **68**(3): 635–640, February 1, 1963.
3. Durney, A. C., Elliot, H., Hynds, R. J., and Quenby, J. J., "Satellite Observations of the Energetic Particle Flux Produced by the High-Altitude Nuclear Explosion of July 9, 1962," *Nature*, **195**(4848): 1245–1247, September 29, 1962.
4. Brown, W. L., and Gabbe, J. D., "Electron Distribution in the Earth's Radiation Belts during July 1962 as Measured by Telstar, *J. Geophys. Res.*, **68**(3): 607–618, February 1, 1963.
5. Hess, W. N., Bloom, S., Mann, L., Seward, F., and West, M., "Electron Loss Rate from the Outer Radiation Belt," *in Space Research III*, ed. by W. Priester, Amsterdam: North-Holland Publishing Co., 1963, pp. 477–485.
6. MacDonald, W. M., and Walt, M., "Distribution Function of Magnetically Confined Electrons in a Scattering Atmosphere," *Ann. Phys.*, **15**(1): 44–62, July 1961.
7. Welch, J. A., Jr., Kaufmann, L., and Hess, W. N., "Trapped Electron Time Histories for $L=1.18$ to $L=1.30$," *J. Geophys. Res.*, **68**(3): 685–700, February 1, 1963.
8. Harris, I., and Priester, W., "Theoretical Models for the Solar-Cycle Variation of the Upper Atmosphere," NASA TN D-1444, August 1962.
9. Van Allen, J. A., Frank, L. A., and O'Brien, B. J., "Satellite Observations of the Artificial Radiation Belt of July 1962," *J. Geophys. Res.*, **68**(3): 619–627, February 1, 1963.
10. McDiarmid, I. B., Burrows, J. R., Budzinski, E. E., and Rose, D. C., "Satellite Measurements in the "Starfish" Artificial Radiation Zone," Div. of Pure Physics, National Research Council, Ottawa, Canada (to be published 1963).
11. Brown, W. L., Gabbe, J. D., and Rosenzweig, W., "Results of the Telstar Radiation Experiments," *Bell Systems Technical Journal*, **42**(4, pt. 2): 1505–1559, July 1963.
12. Welch, J. A., Jr., and Whitaker, W. A., "Theory of Geomagnetically Trapped Electrons from an Artificial Source," *J. Geophys. Res.*, **64**(8): 909–922, August 1959.
13. Parker, E. N., "Effect of Hydromagnetic Waves in a Dipole Field on the Longitudinal Invariant," *J. Geophys. Res.*, **66**(3): 693–708, March 1961.
14. Freden, S. C., and White, R. S., "Particle Fluxes in the Inner Radiation Belt," *J. Geophys. Res.*, **65**(5): 1377–1383, May 1960.
15. Pizzella, G., McIlwain, C. E., and Van Allen, J. A., "Time Variations of Intensity in the Earth's Inner Radiation Zone," *J. Geophys. Res.*, **67**(4): 1235–1254, April 1962.
16. Hess, W. N., "Discussion of Paper by Pizzella, McIlwain and Van Allen 'Time Variations of Intensity in the Earth's Inner Radiation Zone. October 1959 Through December 1960'," *J. Geophys. Res.*, **67**(12): 4886–4887, November 1962.
17. Singer, S. F., "Trapped Albedo Theory of the Radiation Belt," *Phys. Rev. Letters*, **1**(5): 181–183, September 1, 1958.
18. Hess, W. N., Canfield, E. H., and Lingenfelter, R. E., "Cosmic-Ray Neutron Demography," *J. Geophys. Res.*, **66**(3): 665–677, March 1961.
19. McDonald, F. B., and Webber, W. R., "A Study of the Rigidity and Charge Dependence of Primary Cosmic Ray Temporal Variations," *J. Phys. Soc.*, Japan 17 (Suppl. A II): 428–435, January 1962.
20. Brown, W. L., and Gabbe, J. D., "The Electron Distribution in the Earth's Radiation Belts During July 1962 as Measured by Telstar," *J. Geophys. Res.*, **68**(3): 607–618, February 1963.
21. Filz, R., Holeman, E., and Yagoda, H., "Variation of the Directional Flux of Slow Trapped Protons," *AGU Trans.*, **44**(1): 78, March 1963.
22. Garmire, G., "Geomagnetically Trapped Protons with Energies Greater than 350 Mev," *J. Geophys. Res.*, **68**(9): 2627–2638, May 1, 1963.
23. McIlwain, C. E., "Coordinates for Mapping the Distribution of Magnetically Trapped Particles," *J. Geophys. Res.*, **66**(11): 3681–3691, November 1961.
24. Lenchek, A. M., and Singer, S. F., "Geomagnetically Trapped Protons from Cosmic-Ray Albedo Neutrons," *J. Geophys. Res.*, **67**(4): 1263–1287, April 1962.
25. Garren, A., and Ridell, R. J., et al., "Individual Particle Motion and the Effect of Scattering in an Axially Symmetric Magnetic Field," in *Proc. Second U.N. Internat. Conf. on Peaceful Uses of Atomic Energy. Theor. and Exper. Aspects of Controlled Nuclear Fusion, Geneva: United Nations*, p. 65, 1958.
26. Blanchard, R., and Hess, W. N., Private communication, 1963.

NEUTRONS IN SPACE

WILMOT N. HESS

Goddard Space Flight Center

The various neutron sources in the solar system are discussed. Because of the neutrons' half-life, very few neutrons come into the solar system from outside. Cosmic rays interacting with the atmosphere or surface of the planets are the major neutron source near the planets. Studies of the neutron energy spectrum can give some information about the chemical composition of the planets.

The sun is a source of neutrons from several processes. Thermonuclear burn in the corona and leakage from the burning cone do not seem to be important sources, but a significant number of neutrons may result from high energy protons accelerated in flares on the sun.

INTRODUCTION

First, a review of the production process used to explain neutrons in the earth's atmosphere will be given. These neutrons must be produced close by. Since the lifetime of the neutron is only about 10 minutes, neutrons cannot be made outside the solar system and still be observed at the earth. The process that generates neutrons in the atmosphere is the interaction of cosmic ray protons with oxygen and nitrogen nuclei. Neutrons in the energy range from about 0.1 to 5 Mev are made by evaporation, and higher energy neutrons are made by knock-on events. These neutrons slow down by collisions with nuclei, and an equilibrium spectrum is built up.[1] The equilibrium neutron energy spectrum measured in the atmosphere is shown in figure 1. By using neutron diffusion theory, the energy spectrum $\Phi(E)$ at all altitudes has been calculated;[2] and the neutron leakage flux $J(E)$ is obtained by

$$J(E) = D\frac{\partial \Phi(E)}{\partial h} .$$

Summing over energy, we get the expected leakage neutron flux J in space, at geomagnetic latitude λ:

J (neutrons/cm² -sec)	λ (deg)
0.2	0
1.5	90

The energy spectrum of these leakage neutrons at different distances from the earth is shown in figure 2.

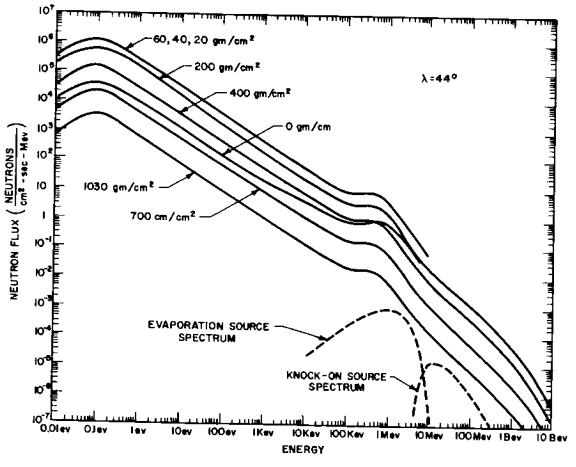


FIGURE 1.—The equilibrium neutron flux versus energy at different depths in the atmosphere for geomagnetic latitude $\lambda = 44°$. The energy spectra for 200 to 1030 gm/cm² are experimental values; for depths less than 200 gm/cm² the spectra are calculated. The shapes of the two neutron source spectra are also shown.

MEASUREMENTS OF THE NEUTRON FLUX

There have been several experiments that measured the neutron flux in space near the earth. A moderated BF_3 counter was flown on an Atlas

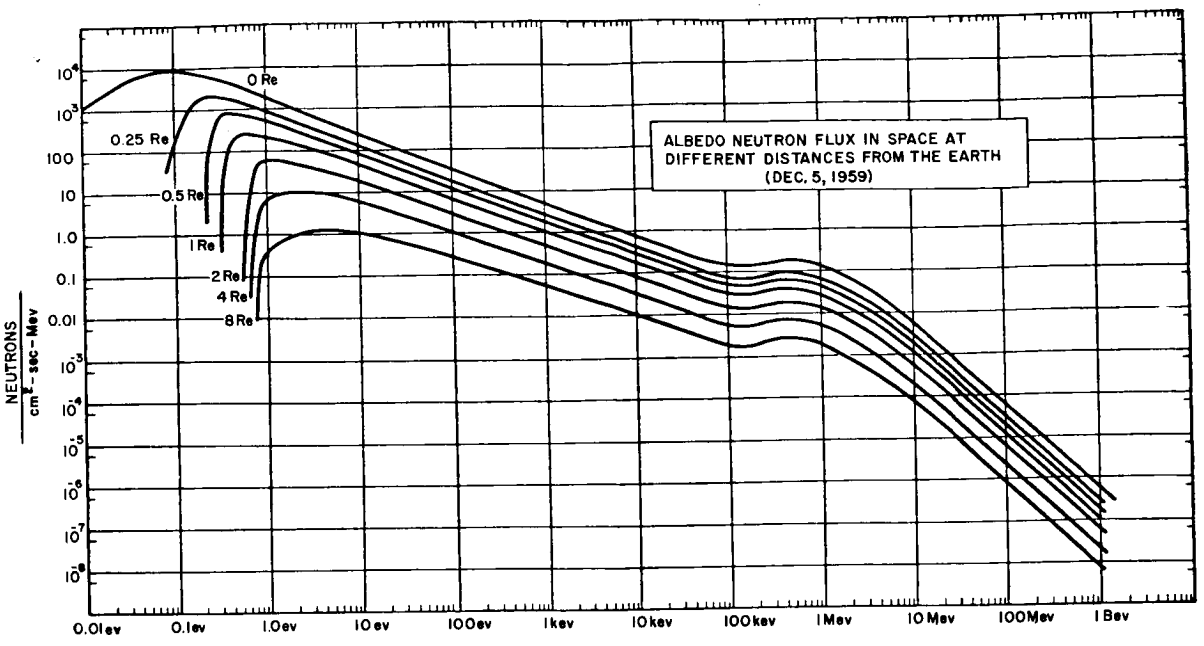


FIGURE 2.—Neutron energy spectra in space at different distances from the earth above the geomagnetic equator (December 1959). The curve for $R_e=0$ is for the top of the atmosphere, which is roughly 100 km.

rocket capsule about two years ago and measured a flux of 1.5 neutrons/cm²-sec at about 40°λ and 200 km altitude[3]—about twice as large as the calculated leakage flux at this location. This difference is probably due to (p, n) events in the capsule. Protons striking the capsule will generate neutrons, some of which will be counted. These give an undesirable background, which needs to be eliminated.

A second BF_3 experiment was performed on an Atlas rocket that flew about three hours after the start on the November 15, 1960, flare. At this time sea level neutron monitors showed a large influx of solar particles. The neutron flux found on this flight was about three times larger than on the first flight and indicated that a new source of neutrons is present—very likely related to the solar proton influx. We will return to this point.

A third BF_3 Atlas experiment, which has not yet been completely analyzed, shows a neutron flux similar to the first flight. This third flight was made during a quiet time, so the flux measured on the first and third flights can be considered a quiet time neutron flux due to galactic cosmic rays.

A recent experiment by Pieper[15] on the TRAAC satellite, using moderated solid state detectors, showed somewhat smaller fluxes than those measured by the BF_3 detectors—smaller by a factor of about 2.

All these measurements are plagued by background troubles. Protons striking the neutron detector will generate neutrons that can be counted with a reasonable efficiency, and protons striking the vehicle carrying the detector will generate neutrons that also can be counted. These two backgrounds are considered separately because they can be eliminated by different experimental techniques. The background due to (p, n) events in the detector has been eliminated by using a moderated neutron detector where the moderator is a scintillator.

A system of this type recently was flown on two Discoverer satellites. An anticoincidence circuit rejects all neutrons made by protons in the detector. A chance coincidence circuit insures that the count rates are not so high that the anticoincidence circuit will not work correctly. The Discoverer satellite flies in a polar orbit at roughly 400 km altitude, so that a complete latitude and longitude survey of the neutron flux

can be obtained in this way. A neutron flux of about 0.3 neutron/cm²-sec was measured at the equator and about 3 neutrons/cm²-sec at the pole. These values are consistent with the Atlas data and are about two times larger than the calculated leakage neutron flux. The problem here is that very likely the background of neutrons made in the vehicle, which are counted by the detector, has not been taken care of or eliminated. Currently two methods are being used to get at this background: calculating it approximately, and also measuring it indirectly. These methods should give a reasonable estimate of the background, but the proper way to deal with it is to eliminate it by placing the detector on an arm far enough away from the vehicle so that the background is negligible. This, coupled with the scintillator moderator, should eliminate all backgrounds. Such detectors will be flown in the future.

NEUTRONS FROM SOLAR PROTONS

Protons coming to the polar caps of the earth from the sun at the times of certain flares provide another source of neutrons.[4] These solar protons will generate neutrons by collisions with O and N nuclei in the same way that galactic cosmic rays do; this can be a quite strong source of neutrons. A large solar proton event might bring 10^5 protons/cm²-sec to the polar caps, which may make 10^4 neutrons/cm²-sec or more leaking out of the polar atmosphere. If this proton event lasts for 10^4 seconds, it will produce about as many neutrons in this short time as galactic cosmic rays do in 1 year.

Two characteristics of these polar neutrons are different from those produced by galactic cosmic rays: They will be of lower energy because the solar protons are mostly of $E<100$ Mev; and, also, they are restricted in their spatial distribution. They will be generated from roughly $\lambda=60°$ to 90°, so they cannot be observed at the equator at low altitude because the earth is in the way. At $\lambda=40°$ at 500 km, where the Atlas measurement was on November 15, 1960, only the edge of the production region can be seen. The enhanced flux observed by this experiment is quite certainly polar neutrons. But, if the experiment had been conducted at $\lambda=60°$ or more, the effect should have been much larger. It is important to try to determine the properties of these polar neutrons accurately because they may play a very important role in the formation of the outer radiation belt. They may be as much as one hundred times more important a source for the outer belt than are the neutrons produced by galactic cosmic rays.[5] Time variations seen in the inner belt, $L\approx 2$, may be due to particles trapped from the decay of polar neutrons.

NEUTRONS FROM THE MOON AND PLANETS

Neutrons will be generated by cosmic rays striking the surface of the moon.[6] This source will be stronger than in the earth's atmosphere for three reasons: First, the moon has a very weak magnetic field, and thus low energy cosmic rays can reach it. Second, the neutron production takes place in a heavier material and, because the cross section for neutron production increases with atomic weight A, more neutrons will be made on the moon. Third, π-mesons made by high energy reactions can interact to make more neutrons on the moon. In the atmosphere of the earth the π-mesons all decay to μ-mesons before interacting, so that no neutrons are made by π-mesons on the earth. The combination of these effects increases the neutron flux by a factor of about 10.

An interesting possibility, related to the leakage neutrons, exists on the moon. The shape of the energy spectrum of the leakage neutrons depends quite sensitively on the amount of hydrogen present, as H is a very powerful neutron moderator. If H is present on the moon, the Mev neutrons are rapidly slowed down to thermal energies and a thermal peak is produced. By measuring the shape of the spectrum, a reasonable estimate of the hydrogen can be made. The presence or absence of H is quite important in understanding the origin and history of the moon.

Galactic cosmic rays make neutrons in the atmosphere of Venus also.[7] Venus' atmosphere is known to be composed of CO_2 with very small amounts of H_2O and CO present. Nitrogen may or may not be there; this is not known. The energy spectrum of the neutrons in the Venusian atmosphere will be changed if N is there. Neither C nor O are good neutron absorbers, but N is; so the population of thermal neutrons will be lessened by the presence of N. This might be one

way of measuring the amount of N$^{\cdot}$ in the Venusian atmosphere.

NEUTRONS FROM THE SUN

The sun will be a source of neutrons from several processes. First, there will be the usual neutrons made by galactic cosmic ray protons. Since the sun has only about 10 percent helium, this neutron source will be considerably weaker for the sun than for the earth. The cosmic ray protons will not generate many neutrons from interacting with the hydrogen in the sun because the threshold for the process

$$p + p \longrightarrow \pi^{+} + n + p$$

is for a proton energy of about 300 Mev. Even above this energy, the cross section for this reaction is small. Thus this cosmic ray neutron source is not important in our considerations.

Another source of neutrons on the sun is thermonuclear reactions in the corona. The corona has a temperature of about 10^6 °K, or the particle energies are about 100 ev. There is a ratio of about 5×10^{-5} of D/H atoms,[8] and we will expect some reactions of the type

$$D + D \longrightarrow n + He^3 .$$

There is so little tritium that reactions involving it are probably not important. The cross section for the DD reaction is dominated by the coulomb barrier penetration term. The cross section times velocity, averaged over a Maxwellian velocity distribution,[9] is given as

$$\langle \sigma v \rangle = 2.6\times10^{-14}\, t^{-2/3} \exp\left[-18.76\, t^{-1/3}\right]$$

$$= .4\times10^{-30} \text{ cm}^3/\text{sec for a temperature of } 10^6\,^{\circ}\text{K}.$$

The reaction note R for this process is given by

$$R = \frac{n^2}{2} \langle \sigma v \rangle \text{ reactions/cm}^3\text{-sec},$$

where n is the number density of D atoms. Assuming that $n=n_0(R/R\odot)^{-m}$, we can write the total production P of neutrons in the corona as

$$P = \int R\, dV = \int k \frac{n_0^2}{2} \left(\frac{R}{R_\odot}\right)^{-2m} \langle \sigma v \rangle 4\pi R^2\, dR \text{ neutrons/sec},$$

where k is the neutron yield per DD reaction. The value of k is about $\frac{1}{2}$ because the probabilities of the two possible reactions,

$$D + D \longrightarrow n + He^3$$

and

$$D + D \longrightarrow p + H^3 ,$$

are essentially equal.

Evaluating $n_0=2\times10^8$ and $m=7.5$ for the known electron densities in the corona,[10] we then get for the neutron flux at the earth:

$$\Phi = \frac{P}{A} = \frac{k n_0^2 \langle \sigma v \rangle 2\pi R_\odot^3}{4\pi R_e^2 (2m-3)} = 5\times10^{-10} \text{ neutron/cm}^2\text{-sec},$$

where R_e is the earth-sun distance. This number is very small and cannot be measured. Reactions below the corona of the sun are not important because the temperature rapidly goes down to 10^4 °K and the reaction rate becomes vanishingly small.

When a big flare comes along, there may be an important number of thermonuclear neutrons in the region of the corona above the flare. The temperature of the corona sometimes increases to about 3×10^6 °K, according to soft x-ray data of Friedman.[11] The D/H ratio can increase to $\frac{1}{10}$, according to Goldberg,[12] at the time of a flare—much larger than the quiet-time value of 5×10^{-5} used here from Severnyi.[8] From these factors,

$$\left(\frac{.1}{3\times10^{-5}}\right)^2 \frac{(.3)^{-2/3}\exp\left[-18.76\,(.3)^{-1/3}\right]}{(.1)^{-2/3}\exp\left[-18.76\,(.1)^{-1/3}\right]} = 1.1\times10^{12}\ .$$

The flare may cover an area of 10^{-4} of the sun's distance, which will give an additional neutron flux at the earth of

$$\Phi = (5\times10^{-10})(1.1\times10^{12})\times10^{-4} = 6\times10^{-2}\ \text{neutron/cm}^2\text{-sec}\ .$$

If the density in a condensation above a flare went up by a factor of 100, the flux would be measurable; but this process does not seem normally to be a big neutron source. There is some evidence for thermonuclear reactions in flares on the sun in the work of Fireman,[13] who analyzed the Discoverer 17 vehicle that flew during a flare and found excess tritium; but, for now, we will pass over this source.

Next we should consider the possibility that neutrons from inside the sun, made in the burning region, can get to the surface. Fowler[14] has given $D/H=10^{-17}$ for the burn region for $T=15\times10^7$ °K. Using the energy output of the sun of 6×10^{10} ergs/cm²-sec, we have

$$\frac{(6\times10^{10})\times4\pi(1.2\times10^{11}\ \text{cm})^2}{28\ \text{Mev}\,(1.6\times10^{-6}\ \text{erg/Mev})} = 2.4\times10^{38}\ \text{pp reactions/sec}\ .$$

Assuming the same cross section for *pp* as for DD, we get about 10^4 neutrons/sec formed. Taking the decay mean life as the lifetime of a neutron, we can obtain an upper limit on the total neutron population at the sum of 10^7 neutrons.

Now, we must see what fraction of these can leak out through about 0.5×10^{11} cm of $\bar{\rho}=1$ material. The important numbers for slowing down and diffusion for H_2O are:

	Length (cm)	Time (sec)
Slowing down	6	10^{-5}
Diffusion	2.9	2×10^{-4}

These numbers are probably rather similar to the sun, which is mostly hydrogen and $\bar{\rho}=1$. From these numbers we see that slowing down to thermal energy occurs very rapidly and then diffusion proceeds. This represents $0.5\times10^{11}/2.9=1.7\times10^{10}$ diffusion lengths, which means $(1.7\times10^{10})^2=3\times10^{20}$ random walk steps, or a time of $(3\times10^{20})\times(2\times10^{-4})$ sec $=6\times10^{16}$ sec. With a factor e of the neutrons lost after each diffusion length, due to absorption, then the fraction of neutrons that leak out of the sun is $e^{-3}\times10^{20}\approx10^{-87}$. This neutron source is rather small!

There is, however, another neutron source on the sun that may not be small. At the time of some large flares, large fluxes of protons arrive at the earth. These fluxes may exceed 10^5 protons/cm²-sec for proton energies of $E_p>10$ Mev, and they may last for $\approx10^4$ sec. This represents at the sun a total energy t of about

$$t = 10^5\ \frac{\text{protons}}{\text{cm}^2\text{-sec}}\times10^4\ \text{sec}\times\left[1/4\pi\times4\pi(1.5\times10^{13}\ \text{cm})^2\right]10^{-14}\ \text{erg/proton}\ ,$$

where we have taken 1 steradian as the size of the proton beam. This gives

$$t = 5 \times 10^{30} \text{ ergs.}$$

This is the order of 1 percent of the observable energy released in a large flare. The mechanics of a flare is not known; but it is not at all impossible that an equivalent- or larger-number of protons go down into the sun and, by interacting, make a large number of neutrons. It is thought that some solar radio bursts may be due to energetic electrons that go down into the sun and make bremsstrahlung. If we guess that an equal number of protons go down and go up, we then have $\approx 5 \times 10^{30}$ protons/sec going downward; and, if they each make on the average 0.01 neutron by interacting with solar material going out in 4π steradians, we will get

$$\frac{(5 \times 10^{30})(.01)}{4\pi (1.5 \times 10^{3})^2} = 26 \text{ neutrons/cm}^2\text{-sec at the earth.}$$

This number will be reduced some by the decay of neutrons on their way to the earth. This is a large flux and should be easily observable at the top of the earth's atmosphere. Because these neutrons will be of low energy $E < 100$ Mev, they will not be seen at sea level but only aloft.

REFERENCES

1. Bethe, H. A., Korff, S. A., and Placzec, G., "On the Interpretation of Neutron Measurements in Cosmic Radiation," *Phys. Rev.*, **57**(7): 573–587, April 1, 1940.
2. Hess, W. N., Canfield, E. H., and Lingenfelter, R. E., "Cosmic-Ray Neutron Demography," *J. Geophys. Res.*, **66**(3): 665–677, March 1961.
3. ———, and Starnes, A. J., "Measurement of the Neutron Flux in Space," *Phys. Rev. Letters*, **5**(2): 48–50, July 15, 1960.
4. Armstrong, A. H., and Harrison, F. B., et al., "Charged Particles in the Inner Van Allen Radiation Belt," *J. Geophys. Res.*, **66**(2): 351–357, February 1961.
5. Lenchek, A. M., "On the Anomalous Component of Low-Energy Geomagnetically Trapped Protons," *J. Geophys. Res.*, **67**(6): 2145–2157, June 1962.
6. Lingenfelter, R. E., Canfield, E. H., and Hess, W. N., "The Lunar Neutron Flux," *J. Geophys. Res.*, **66**(9): 2665–2671, September 1961.
7. ———, Hess, W. N., and Canfield, E. H., "Neutron Leakage Flux and the Nitrogen Abundance of Venus," *J. Atmos. Sci.*, **19**(3): 274–276, May 1962.
8. Severnyi, A. B., "Investigation of Deuterium on the Sun," *Izvestiia Krymskoi Astrofizicheskoi Observatorii*, **16**, 12–44, 1956 (in Russian).
9. Post, R. F., "Controlled Fusion Research—An Application of the Physics of High Temperature Plasmas," *Rev. Mod. Phys.*, **28**(3): 338–362, July 1956.
10. van de Hulst, H. C., "The Chromosphere and the Corona," in: *The Sun*, ed. by G. P. Kuiper, Chicago: University of Chicago Press, 1953, p. 211.
11. Friedman, H., "X-Ray and Ultraviolet Radiation Measurements from Rockets," in: *Space Astrophysics*, ed. by W. Liller, New York: McGraw-Hill, 1961, pp. 107–120.
12. Goldberg, L., Mohler, O. C., and Muller, E. A., "The Profile of $H\alpha$ During the Limb Flare of February 10, 1956," *Astrophys. J.*, **127**(2): 302–307, March 1958.
13. Fireman, E. L., DeFelice, J., and Tilles, D., "Solar Flare Tritium in a Recovered Satellite," *Phys. Rev.*, **123**(6): 1935–1936, September 15, 1961.
14. Fowler, W. A., "Experimental and Theoretical Results on Nuclear Reactions in Stars," *Mémoires de la Société Royale des Sciences de Liège*, Ser. 4, **14**, 88–111, 1954.
15. Pieper, G., (private communication).

ELECTRON LOSS RATE FROM THE OUTER ELECTRON BELT

W. N. HESS

Goddard Space Flight Center

AND

S. D. BLOOM, L. G. MANN, F. D. SEWARD, AND H. I. WEST, JR.

Lawrence Radiation Laboratory

Present experimental information permits a reasonable picture to be formed of the electrons in the outer radiation belt. The lifetime of the trapped particles is very important to a complete understanding of the belts; and for a known steady-state population the only easy way to determine this is to measure the loss rate. This paper discusses the mechanism and rate of the loss of high energy electrons from the outer belt.

Three separate groups of radiation belt electrons have been observed in the Discoverer satellite flights. The first group has a spectrum that falls off sharply with increasing energy, does not extend above 125 kev, and shows large time fluctuations; its source is not understood. The second group has a spectrum similar to that of the auroral electrons. Most of the electrons in the belts are of the third group; their spectrum somewhat resembles a neutron β-decay spectrum but extends to higher energies. They are lost from the belt by coulomb scattering. The calculated loss rate and residence time of these electrons in the outer belt are consistent with the proposition that they are neutron decay products.

INTRODUCTION

Enough experimental information is now available that a resonable picture can be formed of the electrons in the outer Van Allen radiation belt. From Explorer XII(1962υ) we now know that there are typically about 10^7 electrons/cm^2 -sec in the energy region 40 kev to 1 Mev in the outer belt;[1] this is a considerably lower flux than earlier estimates had given. The *flux* does not vary much with position in the outer belt from 25,000 to 65,000 km nor does it vary much with time. *Intensity* variations by a factor of 2 or 3 with time are seen in the 100 kev ebergy range, occasionally reaching a factor of 5 to 7 during August and September (J. J. Van Allen, private communication). These variations might be due only to changes in the magnetic field and the associated effects on particles, rather than to changes in the particle population. The large time variations in flux seen in several earlier experiments are seen only in the high energy ($E>1.6$ Mev) group of electrons in the case of Explorer XII. Very likely the large time variations on several earlier experiments involved the high energy electrons also. At the times of magnetic storms the 100 kev flux changes by as much as a factor of 7, but returns to roughly the prestorm flux in a few days at the end of the storm (J. A. Van Allen, private communication). All of these features seem to indicate a particle population that has quite a long lifetime.

It is very important to a complete understanding of the radiation belt to know the lifetime of the trapped particles. When dealing with a steady-state population, the only easy way to determine particle lifetimes is to measure the loss rate from the belt. The total population of the belt then yields the average lifetime.

In this paper we will consider the mechanism and the rate of the loss of high energy electrons from the outer radiation belt.

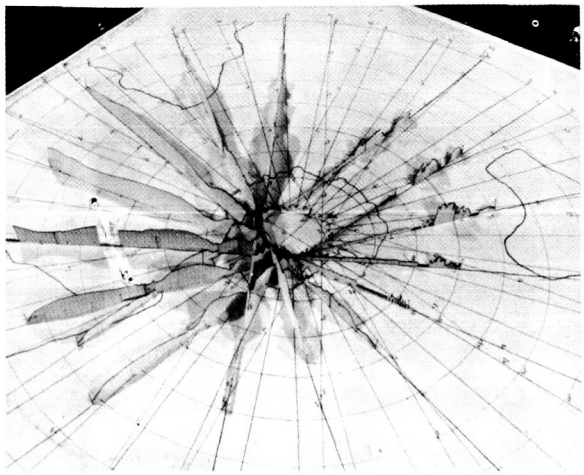


FIGURE 1.—Count rates of the CsI detector on Discoverer 31 in the Southern Hemisphere. The height of the strip gives $\log_{10}$ of the count rate at different positions. The highest count rate is above 10^4 counts/sec and the lowest is about 10 counts/sec.

DATA FROM EXPERIMENTS ON DISCOVERER SATELLITES

We now have information that bears on the lifetime of the outer belt electrons in the 100 kev to 1 Mev energy range. Several particle-detecting instruments have recently flown on Discoverer satellites. These vehicles orbit at altitudes of a few hundred kilometers and at orbital inclinations of approximately 82 degrees, so the earth is quite well covered by measurements from these satellites. The instruments that have been flown are: (1) a thin CsI scintillator covered by 2 mils of Be, which counted electrons of $E\hat{e}>100$ kev and protons above about 3 Mev; (2) a thick plastic scintillator that counted electrons of $E\hat{e}>2$ Mev and protons of $E_p>15$ Mev; and (3) a ten channel magnetic electron spectrometer[2] that counted electrons from 80 kev to 1.25 Mev.

The count rates of the CsI detector are shown in figure 1 for the Southern Hemisphere. Three types of features are seen here. First, off the coast of Brazil a high count rate is observed which is related to the loss of particles from the inner radiation belt. In this region the earth's surface magnetic field is weak, so that the inner belt particles come closest to the earth here. Secondly, irregular pulses of particles are seen in the auroral zones. From the electron spectrometer we know that these particles are electrons of $E\hat{e}<200$ kev and a considerable fraction of them come essentially directly down the field lines. That is, the electrons are not trapped and drifting down in altitude but, rather, they are on their way to earth in one single bounce.

The third group of particles is the most interesting. At about 10 or 15 degrees subauroral in the South Atlantic we see a large population of particles. These particles are not uniquely identified as electrons, but no protons of $E>3$ Mev are known to exist in this region of space, so they are most certainly electrons. These particles show a reasonably consistent spatial distribution from pass to pass. They are the leakage from the outer radiation belt being lost in the Capetown Anomaly.[3] They showed a roll modulation when one of the Discoverer vehicles tumbled, indicating that they are almost certainly trapped particles. These particles were slowly being lost in an orderly way, quite differently from auroral particles which are lost in squirts.

The plastic scintillator shows quite similar behavior in the Southern Hemisphere (figure 2) The losses from the inner belt off Brazil and the loss from the outer belt off Capetown are quite apparent. The auroral spikes are, however, not seen; this indicates that the auroral particles are below the threshold of this detector.

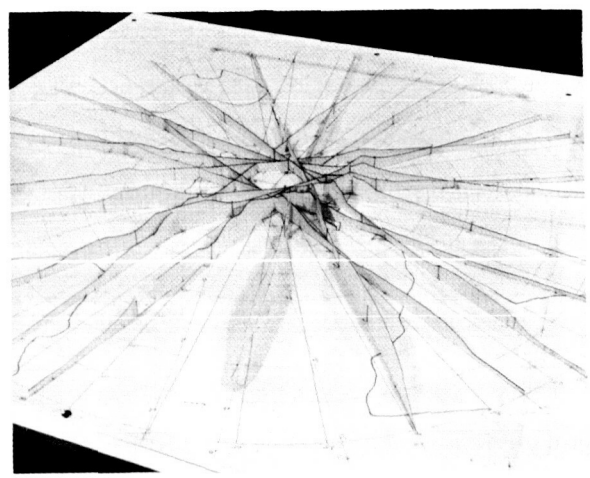


FIGURE 2.—Count rates of the plastic scintillator on Discoverer 31 in the Southern Hemisphere. The height of the strip gives $\log_{10}$ of the count rate at different positions.

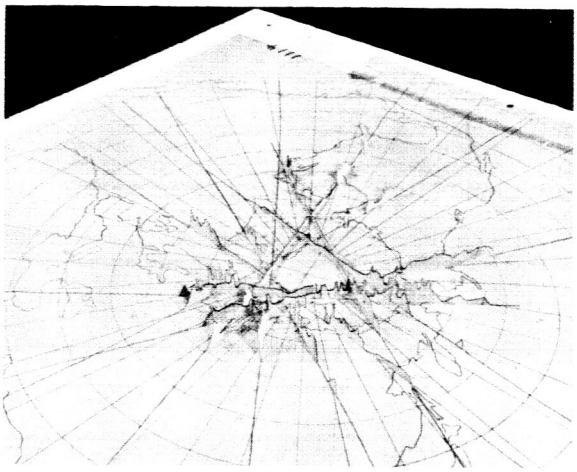


FIGURE 3.—Count rate of the CsI detector on Discoverer 31 in the Northern Hemisphere. The height of the strip gives $\log_{10}$ of the count rate at different positions.

In the Northern Hemisphere the Discoverer results are quite different. Auroral pulses are seen by the CsI detector (figure 3) but the other features are not. The scintillator count rate (see figure 4) shows no high count rate regions such as those in the South Atlantic. Count rate contours for the high count rate regions in the South Atlantic are shown in figures 5 and 6 for the two detectors.

These features of the radiation belts were first

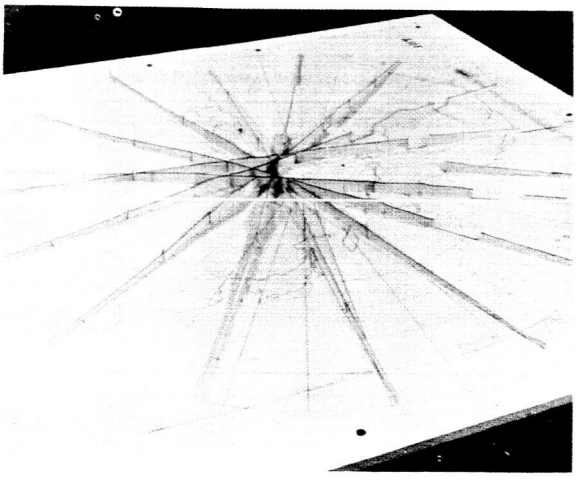


FIGURE 4.—Count rate of the plastic scintillator on Discoverer 31 in the Northern Hemisphere. The height of the strip gives $\log_{10}$ of the count rate at different positions.

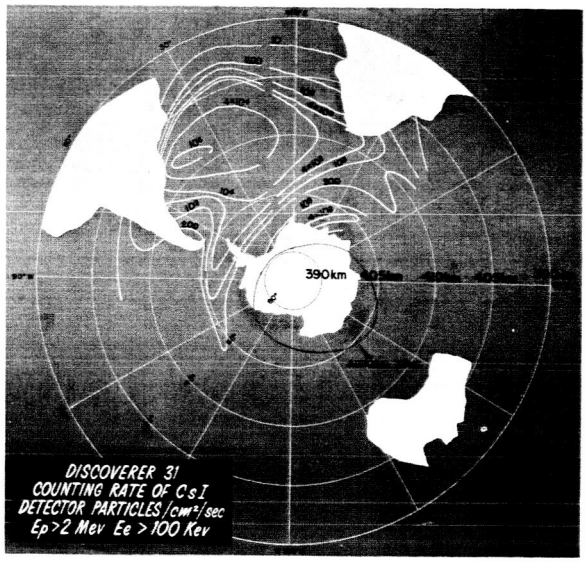


FIGURE 5.—Count rate contours for the plastic scintillator on Discoverer 31 for the Southern Hemisphere.

seen on a low altitude Soviet satellite.[4,5] Vernov et al. observed both the high intensity region off Brazil which they identified as mostly protons and related to the inner belt[4] and a high intensity region in the South Atlantic identified as mostly electrons related to the outer radiation belt.[5]

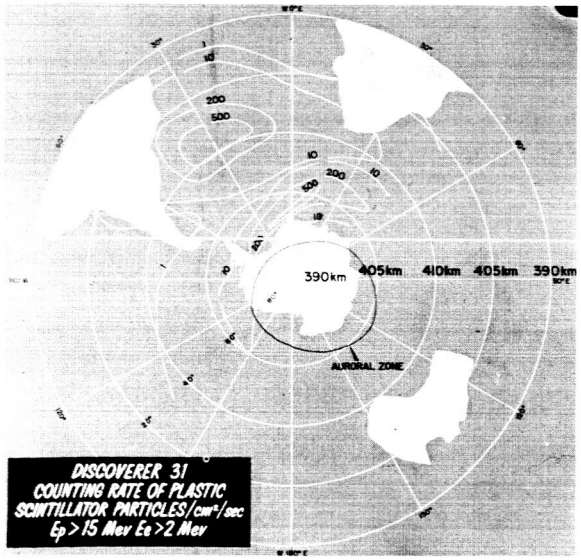


FIGURE 6.—Count rate contours for the CsI detector on Discoverer 31 for the Southern Hemisphere.

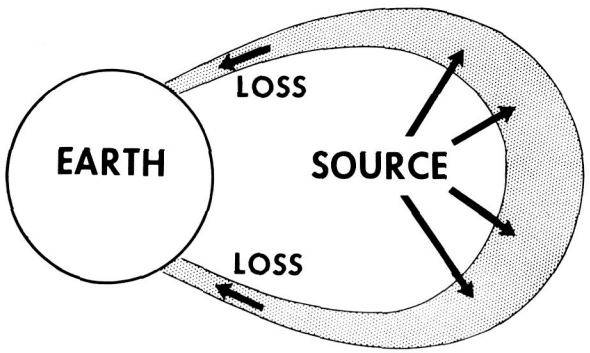


FIGURE 7.—The steady-state picture of the radiation belt. Particles are injected into the belt at a rate s which must equal the loss rate l.

CALCULATION OF THE LOSS RATE

From these low altitude measurements on the outer belt electrons, we can calculate the rate of loss of particles from the radiation belt. Cladis and Dessler[6] suggested studying the bremsstrahlung of electrons in the atmosphere in the region of the Capetown Magnetic Anomaly to get the loss rate from the belt. The present study uses the more direct method of observing the electrons themselves. If we have steady state in the radiation belt, then the situation depicted in figure 7 occurs. Particles are continually injected into the belt. If, as is commonly thought, radial diffusion is a slow process, then the particles are lost from the belt down into the atmosphere in the manner shown in the figure. This loss rate can be described in terms of a motion of particles down into the loss cones and their eventual collision with the earth.

In this calculation of the loss rate, we must assume that coulomb scattering is the dominant loss process. We have measured the particle flux Φ at altitude h. From this we can get the net motion of particles downwards, the drift D, by

$$D = \left(\frac{\Phi}{v}\right) U,$$

where v is the particle's velocity and U is a downward drift velocity. The drift velocity U describes the process of particles gradually leaking out of the loss cone as the result of coulomb scattering. This is, of course, an average concept; a given particle does not move smoothly downwards in altitude. The first attempts to describe this loss process[7,8] introduced the concept of the velocity of lowering of mirror points W. Later it was shown that this description of the process was incomplete[9] and a second term Y was needed which would describe the diffusion of the particles. W. McDonald has shown (private communication) that for the case of the exponential atmosphere the two terms can be written as

$$U = W + Y = W\left[1 + \frac{H}{\Phi\rho}\,\frac{\partial}{\partial h}(\Phi\rho)\right],$$

where H is the scale height of the atmosphere and ρ is the air density.

If the electron flux Φ varies inversely with the air density ρ, then $Y=0$. We know that this situation is roughly true for protons, but there are no good data on electrons to evaluate Y. For lack of better information we will take $W=Y$ and $U=2W$. It would be quite surprising if the altitude dependence of Φ were such that $Y \gg W$, but we must wait for final information on this.

Using the value of $U=1.5\times10^5$ cm/sec for $E=200$ kev at an altitude of 400 km, we can obtain D as a function of position from the CsI detector data. From D we can determine the total loss rate L from the radiation belt by

$$L = \oint D\,da\ .$$

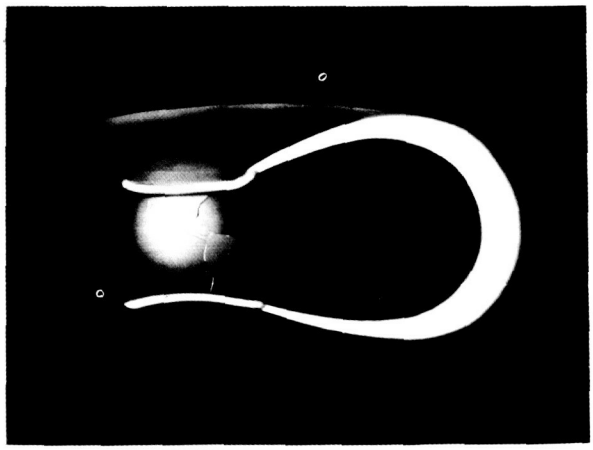


FIGURE 8.—A lunoid of the outer radiation belt. This is a volume whose cross section is the region between two neighboring field lines. The lines on the earth at the base of the lunoid are lines of constant integral invariant which represent the loci of the loss zones for the particles in the lunoid.

To get L we integrate D along two lines of constant integral invariant I, one in the Northern Hemisphere and one in the Southern. This sums up the loss from one lunoid of the belt (a *lunoid* is a figure of revolution whose cross section is a lune, the region between two neighboring field lines, as shown in figure 8). We will take two strips 1 cm wide around the earth in order to evaluate L. The volume of the lunoid whose base is 1 cm wide for $r_0 = 2.8r\hat{e}$ is $V = 1.4 \times 10^{20}$ cm^3. The total loss rate L evaluated this way is 2.5×10^7 electrons/sec and the volume loss rate is $l = L/V = 1.5 \times 10^{-13}$ electrons/cm^3-sec. The only contribution to the integral for L is from the region of the South Atlantic. For a steady-state situation the volume loss rate l is equal to s, the strength of the source that is populating the radiation belt. The source strength s calculated[10] for neutron decay is about 10^{-13} electrons/cm^3-sec for the outer belt. This value of s looks very similar to the calculated value of l; thus the neutron decay source seems adequate in strength to produce the outer belt electrons.

We can also get from this the average residence time τ of an electron in the radiation belt by assuming a steady state and using $L = Q/\tau$, where

$$Q = \left[\frac{10^7 \text{ electrons/cm}^2\text{-sec}}{(2 \times 10^{10} \text{ cm/sec})}\right] 1.4 \times 10^{20} \text{ cm}^3 = 0.7 \times 10^{17} \text{ electrons} .$$

Q is the total number of electrons in the lunoid. We get $\tau = Q/L = 0.7 \times 10^{17}/2.5 \times 10^7 = 3 \times 10^9$ sec. This time τ is about the same as the estimates of lifetimes[11,12,13] for outer belt electrons, based on coulomb scattering.

It is somewhat surprising that the loss rate is as small as it seems to be from this calculation. We might expect, on the basis of other experiments, that polar-cap neutrons would substantially increase the strength of the neutron source due because of galactic cosmic rays. If this were the case, the source strength would be increased above the value of $s = 10^{-13}$.

COMPARISON OF OTHER EXPERIMENTS WITH THE DISCOVERER RESULTS

One other experiment measures the outer belt electron loss rate directly and can be compared with our calculation. The count rate of the 213 GM counter on the Injun satellite[14] determines the loss rate of electrons of $E > 40$ kev. Sometimes the angular distribution of the particles observed by this counter over North Americ is so wide that the particles are obviously being lost directly into the atmosphere. A determination of the loss rate from these observations yields a figure much larger than the one calculated here. According to the Injun analysis, the average residence time τ of an outer belt is about 10^4 seconds. The loss of electrons as observed on Injun must be considered to be the result of a catastrophic process; that is, the electron mirror altitude is changed by a large amount in a single bounce period. This must be due to some process other than coulomb scattering.

This comparison of Injun and Discoverer results seems inconsistent, but with the help of the Discoverer electron spectrometer experiment[2] we can understood both experiments.

The spectrometer detects three different and distinct spectra of elections:

Spectrum A is a very sharply falling off function of energy. The intensity decreases by a factor of e with an increase in energy of about 5 kev, and is only present below 125 kev. This spectrum is seen almost world-wide at some times and other times is not seen at all. At some times, fluxes greater than 10^6 electrons/cm^2-sec-ster are seen in a 28 kev wide channel centered at 94 kev.

Spectrum B e-folds with a change in energy of from 25 to 40 kev and goes up to about 200 kev. It is seen prominently in the auroral zone and shows large time fluctuations. A group of particles of $90 < E < 245$ kev, with a similar energy spectrum and considerably lower intensity, is seen in the inner belt loss zone off Brazil and is seen sometimes in the outer belt loss zone off Capetown.

Spectrum C usually has a maximum intensity at about 600 kev and goes up beyond 1.2 Mev. At

200 kev it is down about a factor of 2 from the maximum intensity. This spectrum is always seen in the inner belt loss zone off Brazil and with lower intensity in the outer belt loss zone off Capetown. The detection of lower intensities of the C spectrum in the outer belt loss zone than off Brazil may be due to the fact that the spectrometer's look direction is closer to the direction of the field line. If the angular distribution is pancake shaped it would be seen less near the direction of the field line. The C spectrum is also seen in certain small areas of the Pacific Ocean. It shows a smooth spatial variation and is quite constant in time. The intensity at one place normally is constant to within a factor of 2 for long periods.

The fact that there are three different spectra of electrons apparently produced by different processes helps explain other experiments. The Injun experiment[14] that observed catastrophic dumping over North America is quite consistent in spatial extent, time variations, energy, and flux with the A spectrum of electrons seen by the Discoverer. The fact that these electrons are almost all of $E<100$ kev means that the CsI detector on Discoverer will not see these particles.

The C spectrum observed by the Discoverer looks somewhat like the equilibrium electron energy spectrum expected[15,16,17] from neutron β-decay, except that it extends to higher energies. This spectrum is rather similar to the outer belt electron seen [18] on Explorer XII in that it is quite flat at about 100 kev and extends up beyond 1 Mev. This population of particles shows smooth spatial variations and roll modulation, indicating that the particles are trapped, and the flux is quite constant in time. All of these facts and also the fact that this spectrum is seen essentially only in the magnetic field anomalies in the South Atlantic strongly indicate that these particles are being lost by coulomb scattering. If other processes were responsible for the particle loss, there would be no reason for the measured flux to be largest in the South Atlantic. The fact that the fluxes of the C spectrum are concentrated in the South Atlantic strongly indicates that the atmosphere controls the loss process. The particles' mirror points come closest to the earth in this region, and the particles encounter the densest atmosphere here. This shows that coulomb scattering is the dominant loss process.

One other experiment has given information on the loss of electrons from the outer radiation belt. Cladia and Dessler[6] analyzed the results of the experiment of Walt et al.[19] who flew a magnetic spectrometer to 1000 km to measure elections. From this analysis a drift flux D of 320 electrons/cm^2-sec was obtained. Assuming the magnetic anomaly is 1000 km wide, this gives a loss rate of $L=3\times10^{10}$ electrons/sec. This is 1000 times larger than the loss rate obtained in the present paper. It is not understood what causes the difference in these values. The electron energy spectrum measured by Walt et al.[19] is not very similar to the C spectrum seen by the Discoverer over the South Atlantic; it more resembles the B spectrum. Possibly the Walt experiment was not seeing trapped outer-belt elections,[18] but some population more like the Injun electrons. This question cannot be answered now. For the present, we must leave this question about the difference of the L values unresolved.

CONCLUSIONS

The following points have come of this analysis:

1. There are three separate groups of electrons observed on the Discoverer flights.

2. The A spectrum falls off sharply with increasing energy and does not extend above 125 kev. It shows large time fluctuations. The Injun experiment that observed direct dumping was quite likely seeing this flux of particles. The source of these particles is not understood.

3. The B spectrum is the spectrum of auroral electrons. A similar spectrum also appears at low energies in some of the particles lost from the radiation belts, indicating that some of the belt electrons of $E<200$ kev may have been accelerated by auroral processes.

4. Most of the electrons in the radiation belt are of the C spectrum type, which looks rather like a neutron β-decay spectrum but extends to higher energies. These particles are lost from the radiation belt by coulomb scattering.

5. The average residence time of an electron in the outer radiation belt is calculated on the basis

of coulomb scattering to be 3×10^9 seconds. The analysis here is uncertain by a factor of 2 or more in a few instances, but the results are probably good to within a factor of 5.

6. The data on the C spectrum electrons concerning their energy spectrum, loss rate, and lifetime are all consistent with the proposition that these particles, which constitute most of the inner and outer radiation belt electrons, are the result of neutron decay.

REFERENCES

1. ROSSER, W. G. V., and O'BRIEN, B. J., et al., "Electrons in the Earth's Outer Radiation Zone," paper presented at 43rd Annual Meeting Amer. Geophys. Union, Washington, April 1962.
2. MANN, L. G., BLOOM, S. D., and WEST, H. I., Jr., "The Electron Spectrum from 80 to 1258 Kev Observed on Discoverer Satellites 29 and 31," paper presented at the 3rd Internat. Space Sci. Sympos., Washington, May 1962.
3. DESSLER, A. J., "Effect of Magnetic Anomaly on Particle Radiation Trapped in Geomagnetic Field," *J. Geophys. Res.*, **64**(7): 713–715, July 1959.
4. VERNOV, S. N., and SAVENKO, I. A., et al., "Discovery of the Internal Radiation Belt at a Height of 320 km in the Region of the South Atlantic Magnetic Anomaly," *Doklady Akademii Nauk SSSR*, **140**(5): 1041–1044, October 11, 1961. *Translation in Soviet Phys.—Doklady*, **6**(10): 893–896, April 1962.
5. ———, ———, and SHAVRIN, P. I., "The External Radiation Belt About The Earth at an Altitude of 320 Kilometers," *Doklady Akademii Nauk SSSR*, **140**(4): 787–790, October 1961.
6. CLADIS, J. B., and DESSLER, A. J., "X-Rays from Van Allen Belt Electrons," *J. Geophys. Res.*, **66**(2): 343–350, February 1961.
7. CHRISTOFILOS, N. C., "Trapping and Lifetime of Charged Particles in the Geomagnetic Field," Univ. Calif. Radiation Lab. Rept. UCRL–5407, November 28, 1958.
8. WELCH, J. A., and WHITAKER, W. A., "Theory of Geomagnetically Trapped Electrons from an Artificial Source," *J. Geophys. Res.*, **64**(8): 909–922, August 1959.
9. KELLOGG, P. J., "Possible Explanation of the Radiation observed by Van Allen at High Altitudes in Satellites," *Nuovo Cimento*, **11W**, 48–66, January 1, 1959.
10. HESS, W. N., CANFIELD, E. H., and LINGENFELTER, R. E., "Cosmic-Ray Neutron Demography," *J. Geophys. Res.*, **66**(3): 665–677, March 1961.
11. WENTWORTH, R. C., "Lifetimes of Geomagnetically Trapped Particles Determined by Coulomb Scattering," Ph.D. Thesis, University of Maryland, 1960.
12. ———, MACDONALD, W. W., and SINGER, S. F., "Lifetimes of Trapped Radiation Belt Particles Determined by Coulomb Scattering," *Physics of Fluids*, **2**(5): 499–509, September–October 1959.
13. HESS, W. N., and KILLEEN, J., "Spatial Distribution of Electrons from Neutron Decay in the Outer Radiation Belt," *J. Geophys. Res.*, **66**(11): 3671–3680, November 1961.
14. O'BRIEN, B. J., "Direct Observations of Dumping of Electrons at 1000 km Altitude and High Altitudes," State Univ. of Iowa, SUI–62–2, January 10, 1962.
15. KELLOGG, P. J., "Electrons of the Van Allen Radiation, *J. Geophys. Res.*, **65**(9): 2705–2713, September 1960.
16. LENCHEK, A. M., SINGER, S. F., and WENTWORTH, R. C., "Geomagnetically Trapped Electrons from Cosmic Ray Albedo Neutrons," *J. Geophys. Res.*, **66**(12): 4027–4046, December 1961.
17. HESS, W. N., and POIRIER, J. A., "Energy Spectrum of Electrons in the Outer Radiation Belt," *J. Geophys. Res.*, **67**(5): 1699–1709, May 1962.
18. O'BRIEN, B. J., and VAN ALLEN, J. A., et al., "Absolute Electron Intensities in the Heart of the Earth's Outer Radiation Zone," *J. Geophys. Res.*, **67**(1): 397–402, January 1962.
19. WALT, M., CHASE, L. F., CLADIS, J. B., IMHOF, W. L., and KNECHT, D. J., "Energy Spectra and Altitude Dependence of Electrons Trapped in the Earth's Magnetic Field," in *Space Research I*, ed. by H. Kallman-Bijl, p. 910–920.

ENERGY SPECTRUM OF ELECTRONS IN THE OUTER RADIATION BELT

WILMOT N. HESS AND JOHN A. POIRIER†

Goddard Space Flight Center

The equilibrium energy spectrum of electrons in the outer radiation belt is determined by the injection spectrum and the loss processes that operate to remove the electrons or change their energy. The loss processes considered here are ionization energy loss, multiple scattering, and electron-electron scattering; the injection spectra considered are neutron β-decay electrons and monoenergetic electrons of 780 and 20 kev. The problem is treated numerically. The results of the numerical calculation are compared with recent measurements of the outer-belt electron spectrum; it appears that neutron decays produce a reasonable fraction of the outer-belt electrons, but other processes such as acceleration may be important.

INTRODUCTION

The sources of particles that populate the Van Allen radiation belts present an interesting problem. It is quite well established now that most of the protons in the inner belt come from the decay of neutrons leaking out of the earth's atmosphere.[1,2] The source of electrons is not so well determined. It would be surprising if the electrons in the inner radiation belt were not also from leakage neutron decay,[3] but this is not certain.

We will examine here the kind of an electron belt that would be produced at 2 or 3 earth radii out from the surface of the earth by neutron decay alone. We have detailed quantitative knowledge of the neutron-decay source strength[4] and therefore can quantitatively study the equilibrium electron spectrum to be expected in this region. Then, by comparison with experimental information about the outer radiation belt, we can see whether neutrons produce an important fraction of the outer belt.

We know the neutron source strength adequately but, to determine the equilibrium electron spectrum, we must understand the loss mechanisms operating to remove electrons. One such process is multiple small-angle Coulomb scattering,[5,6,7] which changes the pitch angle of the particles and lowers the mirror point so that the particles get lost in the atmosphere. Another process is ionization energy loss, which degrades the electron's energy. The former is proportional to Z^2, where Z is the atomic number of the gas present; the latter is proportional to Z. Therefore the dominant loss mechanism depends on the type of material present.

Particles in the inner radiation belt move most of the time in an atmosphere of oxygen and nitrogen. The higher Z of this material is such that multiple scattering is the principal loss mechanism. Particles in the outer belt exist essentially in a hydrogen exosphere, where ionization energy loss is important.

A typical high-energy electron in the outer radiation belt might move on a magnetic line of force going out 24,200 km from the center of the earth at the equator. Let us consider a 780-kev electron that has a mirror point at 30 degrees magnetic latitude. Let λ be the magnetic latitude of a field line, R the distance from the center of the earth, and α the pitch angle of the electron spiraling about the magnetic field line. Tan$\alpha = v_\perp/v_{||}$, where $v_\perp$ and $v_{||}$ are velocity components perpendicular and parallel to the magnetic field line respectively. Let us approximate the earth's

†Lawrence Radiation Laboratory, Univ. of California, Berkeley, California.

TABLE 1.—*Values of α and λ Satisfying Equations 1, 2, and 3 for a Mirror Point at (a) 17,200 km and (b) 8000 km*

R (km)	λ (deg)	α (deg)	
		Case (a)	Case (b)
24,200	0	34.4	8.3
17,200	30	90	14.8
8,000	54.9	—	90

magnetic field by a dipole field for which the field lines are given by

$$R = R_0 \cos^2 \lambda , \quad (1)$$

where R_0 is the radius of the field line at the equator ($\lambda = 0$). Its strength is given by

$$B = CR^{-3} \sqrt{1 + 3\sin^2 \lambda} , \quad (2)$$

where $C = 8.1 \times 10^{10}$ gauss-km^3. The pitch angle of the electron must satisfy the relation

$$B = B_m \sin^2 \alpha , \quad (3)$$

where B_m is the value of the magnetic field at the mirror point. Table 1 gives values of α and λ that satisfy equations 1, 2, and 3 for: (a) a mirror point at 17,200 km, and (b) a mirror point at 8000 km.

Essentially, to lose a particle from the radiation belt by scattering, its angle must be altered so that the mirror point is lowered to about 8000 km, where it will encounter a relatively dense atmosphere. Table 1 shows that, if the scattering takes place at the equator, a change of pitch angle of 26.1 degrees is necessary to change the mirror point from 17,200 to 8000 km and, if the scattering takes place at 17,200 km, a change of 75.2 degrees is necessary. As an average intermediate situation, we will consider that a change of pitch angle of 1 radian will remove a particle from the outer belt. This corresponds to a scattering through a projected scattering angle of 1 radian, or a total scattering angle of $\sqrt{2}$ radians.

To determine which loss mechanisms are dominant in the outer belt, we must consider a

TABLE 2.—*Summary of Scattering Probabilities as a Function of Particle Range*

E (kev)	R (gm/cm^2)	ΔR (gm/cm^2)	$\bar{E}$ (kev)	$\overline{\theta_P^2}$ (rad^2)	$\Sigma\overline{\theta_P^2}$ (rad^2)	ΣP_{MS}	ΣP_{SS}	ΣN_{ee}
780	0.153	0.020	740	0.016	0.016	0.0	0.0005	0.11
700	0.133	0.025	650	0.026	0.042	0.0	0.0012	0.24
600	0.108	0.024	550	0.033	0.075	0.0003	0.0023	0.37
500	0.084	0.023	450	0.045	0.12	0.0031	0.0037	0.50
400	0.061	0.0213	350	0.065	0.19	0.024	0.0059	0.63
300	0.0397	0.0187	250	0.10	0.29	0.064	0.0097	0.74
200	0.0210	0.0144	150	0.19	0.48	0.16	0.018	0.83
100	0.00657	0.00463	75	0.20	0.68	0.22	0.029	0.85
50	0.00194							

change of pitch angle by Coulomb scattering and energy loss by ionization. For particles traveling in unionized hydrogen gas, Table 2 lists the probabilities of multiple scattering and single scattering to projected angles greater than 1 radian, and scattering from an electron that produces a recoil electron of greater than 30 kev energy. This calculation is summarized in Table 2 as a function of the particle's range. In Table 2, E is the electron kinetic energy in kev, R is its range in hydrogen in gm/cm^2 (Ref. 8), ΔR is the amount of hydrogen necessary to lower the electron energy from the top of the bottom of the rnergy group indicated, $\bar{E}$ is the average kinetic energy of the group in kev, $\bar{\theta}_P{}^2$ is the square of the rms projected multiple-scattering angle in rad^2 (Ref. 9), ${}_8\sum\bar{\theta}_P{}^2$ is the sum of these angles to get the additive effect, ${}_8\sum P_{MS}$ is the probability that this value of ${}_8\sum\bar{\theta}_P{}^2$ will scatter the particle downward 1 radian or more, $\sum P_{SS}$ is the probability that the electron will single-scatter 1 radian or more in the projected angle downwards, and $\sum N_{ee}$ is the number of recoil electrons from electron-electron collisions having energy 30 kev or higher.[9]

Inspection of this table shows that an electron will slow down from 780 to about 50 kev before the scattering gets large enough that a significant number of particles will be lost from the outer belt ($\sum P_{MS}=0.22$ at 50 kev). Below 50 kev, the scattering increases sharply, and a substantial fraction of the particles will be lost by scattering. Only about a quarter of the particles in the outer belt will scatter out of the belt when a 780-kev electron slows down to 50 kev. About one secondary electron of energy greater than 30 kev will be made per 780-kev electron slowing down. Thus in the numerical calculation to follow we must consider both slowing down and scattering as important loss mechanisms in the outer zone, with the additional production of low-energy electrons by electron-electron collisions.

NUMERICAL CALCULATIONS

The scattering calculations were coded for an IBM 709 computer. The program was run with several input spectra. The first problem was started by having all particles in one energy group (770 to 778 kev) and was stopped when all particles were slowed down below 30 kev. This case of a monoenergetic pulse source can be checked analytically, and it demonstrates the kinds of physical processes that take place. This problem is moderately similar to what happened in the Argus experiment when a pulse of electrons was suddenly injected into the earth's magnetic field by a nuclear explosion.[10]

The second calculation starts with the neutron β-decay spectrum and follows these particles down to 30 kev. This problen has a bearing on the makeup of the other radiation belt.

Each of these calculations have been studied for: (1) a pulse source, and (2) a continuous source. In case (1) a squirt of particles is started all at one time and slowed down together. In case (2) the same number of particles are injected at each interval in time, and the problem is carried forward until an equilibrium spectrum is obtained.

The computer program starts with an initial spectrum and calculates the spectrum after a short time Δt. The time step is taken short enough that the probability of a particle with $E=30$ kev scattering twice in one time step is very small. The process is repeated for many time steps until the problem is completed. Three loss mechanisms have been considered: (a) the usual ionization energy loss for a moving charged particle having distant collisions with electrons (dE/dx); (b) close collisions with electrons, which occur less frequently but are important in that larger momentum transfers are involved and new, fast particles can be made in this way; and (c) the loss by scattering, which removes particles from the region by changing the pitch angle.

Loss processes (a) and (b) are handled together, while (c) is considered separately. To handle (c), we calculate how many particles $S(E)$ are lost from each energy group in each time step by scattering. The fraction of particles scattering in each step is obtained from

$$S(E) = \frac{\Delta t\bar{\rho}}{[\rho\tau(E)]}, \qquad (4)$$

where $\bar{\rho}$ is the number of protons per cubic centimeter traversed by the electron and $\bar{\rho}\tau(E)$ is the product of proton density and electron lifetime as calculated by Wentworth, MacDonald, and Singer.[7]

The energy loss processes (a) and (b) have

been represented by

$$N^i(m+1) = N^i(m) + \sum_{l=i+1}^{l_{max}} N^l(m)\,P_i^{\,l} - N^i(m) \sum_{l=i/2}^{i-1} P_l^{\,i}$$

$$- N^i(m)\left|\frac{dE^i}{dx}\right|_c \frac{\Delta x^i}{\Delta E} + N^{i+1}(m)\left|\frac{dE^{i+1}}{dx}\right|_c \frac{\Delta x^{i+1}}{\Delta E} , \tag{5}$$

which relates the number of particles in the ith energy group at time step $m+1$, $N^i(m+1)$, to those in that group at time step m, $N^i(m)$. Here $P_i^{\,l}$ is the probability that an electron be scattered from group l (superscript) to group i (subscript) by an electron-electron scatter. The first term on the right side of equation 5 is the number of particles in group i just before time step $m+1$. The second term is the number that scatter into group i from energies above it; this electron may be a recoil electron or a higher energy electron that has been degraded in energy (these two electrons are indistinguishable). The third term is the number of particles that scatter out of group i into lower energy groups. The fourth term is the number of particles that lose enough energy by ionization to be degraded from the ith group to the next lower energy group. The last term is the number of particles in the $i=1$ group that lose enough energy to be degraded into the ith group.

The probability $P_i^{\,l}$ is

$$P_i^{\,l} = (v\rho_e \Delta t)\left(\frac{d\sigma_{ee}}{dW}\right)\Delta E^i , \tag{6}$$

where ρ_e is the electron density. In equation 6 the electron-electron scattering cross section has been calculated by C. Møller:[11]

$$\frac{d\sigma_{ee}}{dW} = \frac{2\pi r_0^2}{\beta^2}(mc^2)\left[\frac{1}{W^2} + \frac{1}{(E-W)^2} + \frac{1}{(E+mc^2)^2} - \frac{2E+mc^2}{(E+mc^2)^2}\frac{1}{W(E-W)}\right] , \tag{7}$$

where $2\pi r_0^2 = 2\pi(e^2/mc^2)^2 = 0.50\times10^{-24}$ cm^2; $\beta = v/c$ for the incident electron; m is the rest-mass energy of the electron; E is the incident-electron kinetic energy; and W is the kinetic energy of either of the two final electrons. All quantities are expressed in the laboratory coordinate system.

The ordinary ionization energy loss corrected for that part of the energy loss that has already been included as large-angle scattering in the second and third terms in equation 5 is

$$\left|\frac{dE^i}{dx}\right|_c = \left|\frac{dE^i}{dx}\right| - \sum_{l=i/2}^{i-1} \frac{d^2P_l^{\,i}}{dE\,dx}\left(E^i - E^l\right)\Delta E , \tag{8}$$

where[12]

$$\left|\frac{dE}{dx}\right| = \rho_e mc^2 \frac{2\pi r_0^2}{\beta^2}\left[\log_e \frac{mc^2 E\beta^2}{2I^2(1-\beta^2)}\right.$$

$$\left. -\left(2\sqrt{1-\beta^2} - 1 + \beta^2\right)\log_e 2 + 1 - \beta^2 + \frac{1}{8}\left(1 - \sqrt{1-\beta^2}\right)^2\right] . \tag{9}$$

Here ρ_e is the number of electrons per cubic centimeter, I is the ionization energy constant, and the

rest of the symbols are as defined above. Combining equations 5 and 8, we get

$$N^i(m+1) = N^i(m) + \sum_{l=i+1}^{l_{max}} N^l(m) P_i^l + N^i(m) \sum_{l=i/2}^{i-1} (i-l-1) P_l^i$$

$$+ N^{i+1}(m) \left|\frac{dE^{i+1}}{dx}\right| \frac{\beta^{i+1} c \Delta t}{\Delta E} - N^i(m) \left|\frac{dE^i}{dx}\right| \frac{\beta^i c \Delta t}{\Delta E}$$

$$- N^{i+1}(m) \sum_{l=(i+1)/2}^{l_{max}} (i-l+1) P_l^{i+1} . \qquad (10)$$

In the energy region from 30 to 60 Mev, additional approximations have improved the convergence of the sums. The character of the spectrum on either side of 60 kev lends validity to these approximations.

The calculation for each time step is repeated until all particles have slowed down below 30 kev. The time interval Δt has been selected to be small enough so that there is a very small chance that a particle will be scattered twice within that time step. Since the chance of scattering goes up with a decrease in particle energy, a compromise must be made between computer time and the lowest energy of interest. On this basis, the lowest energy of interest was chosen to be 30 kev.

RESULTS

As was stated before, in the first problem we started with a pulse of particles at $E=780$ kev and followed them as they produced secondary electrons, were slowed down by ionization energy loss, and were scattered out. The starting datum of the second problem was the neutron-decay spectrum.

The 780-kev Electron Problem

The first problem was undertaken to demonstrate the physics in the slowing-down problem. Starting with all the particles at 780 kev, the calculation proceeds until all the particles slow down to less than 30 kev. Figure 1 shows several steps in the slowing-down process. In a short time, the initial pulse of particles, point A in figure 1, turns into a U-shaped distribution, curve B, by producing knock-on electrons. Curves B, C, D, and E are drawn for equal time intervals where the energy loss by ionization as well as the buildup of low-energy recoil electrons can easily be seen. Curve C′ (dashed line) is the

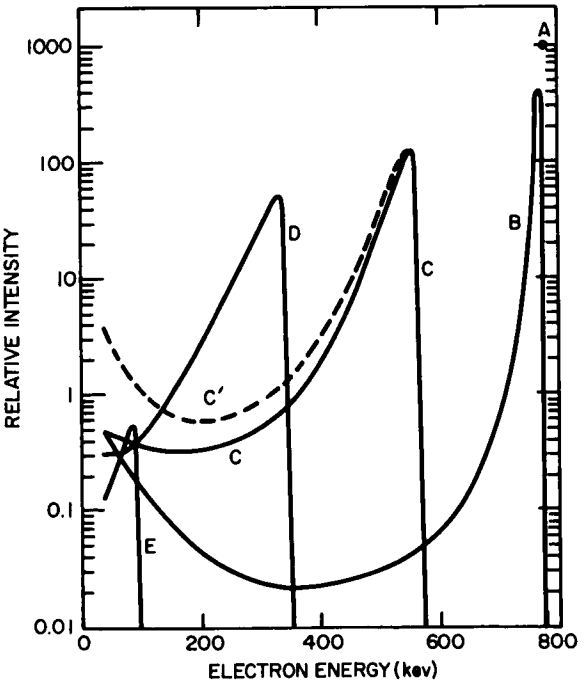


FIGURE 1.—A single group of electrons injected at 780 Kev (point A) are slowed down, produce knock-on electrons, and scatter out of the region of interest in the outer belt. Curve B shows a time soon after the injection time; curves C, D, and E are at equal time intervals after injection. The dotted part of curve C shows the effect of not including the Coulomb scattering-out term in the calculation.

solution for the problem ignoring the scattering-out term. We see that the effect of this scattering-out term becomes large only for energies below about 100 kev. Eventually all the particles are lost by scattering or pass below the energy cutoff of the problem. It should be pointed out that these figures give the density of particles at a given energy (electrons/cm^3). If a flux of particles is desired (electrons/cm^2-sec), the spectra must be multiplied by the appropriate velocity.

The Argus nuclear explosion,[13] carried out at about 300 miles above the earth, formed a shell of trapped electrons in space. The electrons were produced by β-decay of the fission fragments from the explosion. These electrons were lost eventually by multiple small-angle Coulomb scattering[6] on the oxygen and nitrogen atoms in the atmosphere. However, if the Argus explosions had been at an altitude of 2000 km or more and had therefore been in the hydrogen exosphere, the electron spectrum would have behaved roughly as is shown in figure 1 because of the three loss mechanisms considered. There is, of course, a spectrum of energy for electrons born of a nuclear explosion, some of which have energies of more than 1 Mev. The physical processes remain the same however, and the above solution for a δ-function input should give a qualitative feeling for what would be expected for a high-altitude Argus experiment.

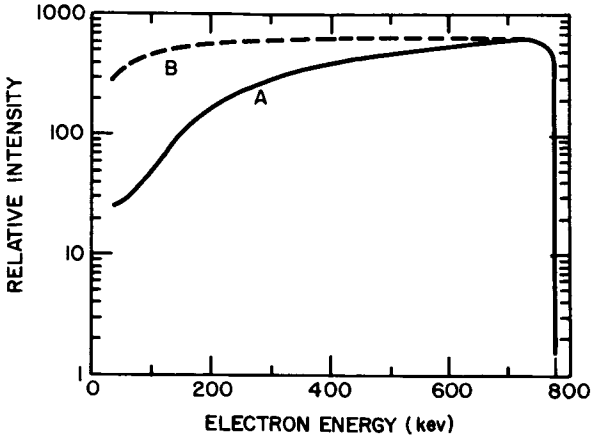


FIGURE 2.—Curve A is the equilibrium spectrum produced by a continuous injection of electrons of 780-Kev energy into the outer radiation belt. The dashed curve B is the equilibrium spectrum if we assume that ionization energy loss is the sole loss mechanism.

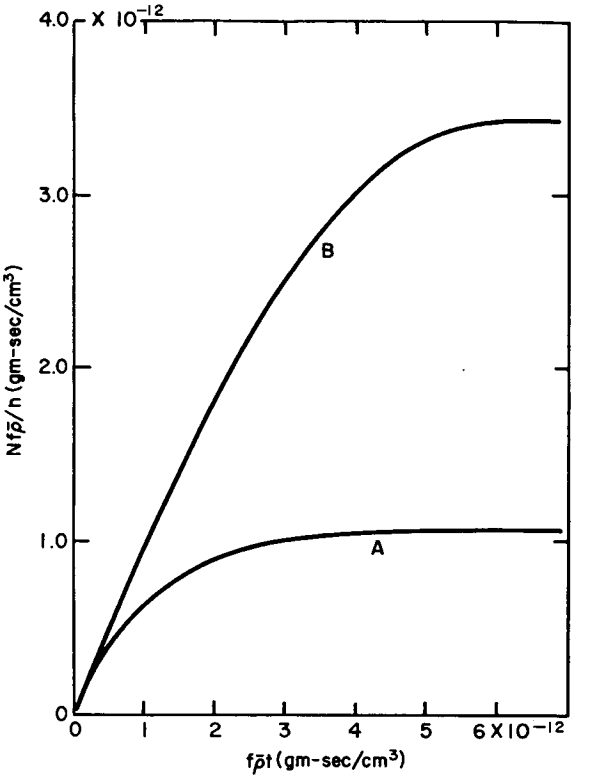


FIGURE 3.—The buildup toward equilibrium by a continuous injection of electrons in the outer belt. Curve A is, in units defined below, the total number of electrons (above 30 Kev) as a function of time after the start of uniform injection of electrons with a β-decay energy spectrum. Curve B is the same quantity for a δ-function source of electrons at 780-Kev kinetic energy. On the abscissa the quantity $f\bar{\rho}t$ is plotted in gm-sec/cm^3, where t is the time in seconds, $\bar{\rho}$ is the average density of hydrogen in gm/cm^3, and f is a factor that describes the increased effectiveness of a partially ionized medium to multiple-scatter and absorb energy by ionization energy loss (see text). The quantity $Nf\bar{\rho}/n$ is plotted on the ordinate, where N is the number of electrons/cm^3 at equilibrium and n is the number of electrons/cm^3-sec that are injected.

If we consider a situation where a stream of 780-kev electrons is continuously introduced into the magnetic field, an equilibrium spectrum will result. The solid line A in figure 2 shows the spectrum that would result after equilibrium is established. The total number of particles as a function of the time after initiating the injection (the buildup toward equilibrium) is given in figure 3, curve B. An example of this physical process might be a particle accelerator in space

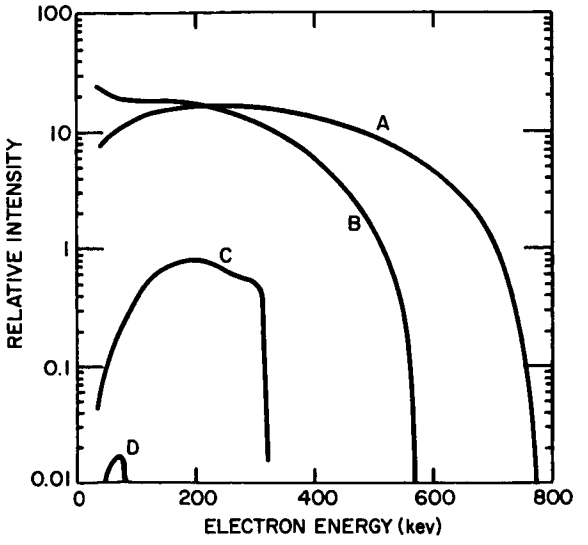


FIGURE 4.—Curve A is the neutron β-decay energy spectrum. Curves B, C, and D show the changes that would occur at equal time intervals after a single burst of β-decay electrons.

continuously emitting a beam of 780-kev electrons.[13]

Several checks have been made to insure that the computer program was doing the desired calculation. One such check has been indicated in figure 2 by a dashed line, curve B. This line indicates the equilibrium spectra that would be attained if there were a continuous source of 780-kev electrons and the only loss mechanism was ionization energy loss. This case can easily be calculated analytically, and the results agree with the machine calculations. To illustrate the nature of the analytic solution, first assume that dE/dx (ionization energy loss per centimeter of stopping material) has the simplified form

$$\frac{dE}{dx} \approx kv^{-2} , \qquad (11)$$

where k is a constant and v the electron velocity. The probability dP/dE of finding a particle between energy E and dE is directly proportional to the time dt that the particle spends in this energy interval. Therefore we can write

$$\frac{dP(E)}{dE} = \frac{dt}{dE} = \left[\left|\frac{dE}{dx}\right|\frac{dx}{dt}\right]^{-1} \approx \frac{v}{k} = \frac{1}{k}\frac{(E^2 + 2Em)^{1/2}}{E + m}, \qquad (12)$$

where m is the mass and E the kinetic energy of the electron. Starting with the relativistic dE/dx formula, equation 9, for the ionization energy loss of electrons, a correct relativistic calculation of dP/dE has been made following the above outline. Results of this analytic calculation agree very closely with those of the corresponding machine calculation, which is shown by curve B in figure 2.

Neutron β-Decay Electrons

The main purpose of this paper is to calculate what electron spectrum would exist in the outer radiation belt resulting from neutron β-decay electrons alone. Starting with a pulse of neutron β-decay electrons (curve A, figure 4), the calculations proceed as before. The slowing down and scattering out is shown in several stages in figure 4. Curves B, C, and D are drawn for equal time intervals.

The equilibrium spectrum (curve A, figure 5) is compared with the original β-decay spectra, curve B in figure 5. Curve A in figure 5 is the electron energy spectrum that would exist in the outer radiation belt if neutron decay was the only source of electrons and we had considered all the loss processes. These data are also electron density rather than flux. We see that the equilibrium spectra is much the same shape as the

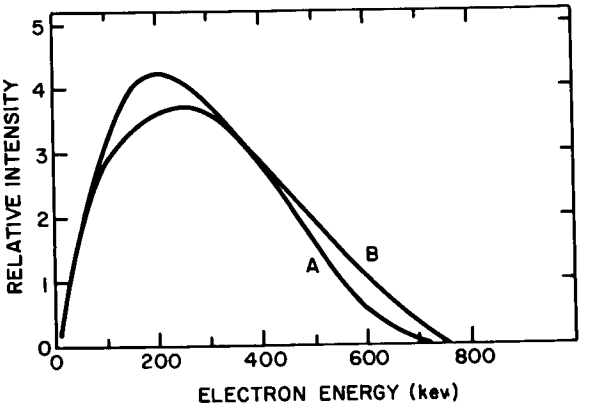


FIGURE 5.—Curve B is the electron energy spectrum from neutron β-decay. With this as the energy spectrum of the electrons injected into the outer belt, the equilibrium energy distribution of curve A was obtained after considering the loss processes discussed in the text. The plot is electron density (not flux) per Kev on a linear scale.

original β-decay spectrum except for a slight shift toward lower energy. The buildup toward equilibrium as a function of time is shown in figure 3, curve A. This figure gives the total number N of particles with energies greater than 30 kev as a function of time t. The units on the abscissa are $f\bar{\rho}t$, where f is a dimensionless number that characterizes the increased effectiveness of a partially ionized medium to multiple-scatter or to lose energy by ionization energy loss, $\bar{\rho}$ is the average density in gm/cm^3, and t is the time in seconds. The units on the ordinate are $f\bar{\rho}N/n$, where N is the number of particles/cm^3 in the radiation belt and n is the rate of injection in electrons/cm^3-sec.

For example, assume that a particle mirrors at 30 degrees magnetic latitude on the line that reaches a maximum distance of 3.8 earth radii from the center of the earth. Assume that the hydrogen is 50 percent ionized and has an average density of 10^{-21} gm/cm^3.[14,15] From figure 3 we see that the electron spectra reaches 50 percent of its equilibrium value for $f\bar{\rho}t=0.8\times10^{-12}$ gm-sec/cm^3. Here f is

$$f = \frac{\left|\frac{dE}{dx}\right|_{ion} f_{ion} + \left|\frac{dE}{dx}\right|_{neut} f_{neut}}{\left|\frac{dE}{dx}\right|_{neut}}, \quad (13)$$

where $|\,dE/dx\,|_{ion}$ is the ionization energy loss by completely ionized hydrogen and f_{ion} is the fraction of the hydrogen that is ionized. The subscript "neut" denotes the corresponding quantities for neutral hydrogen. The same kind of an equation can be written for the multiple-scattering angle.[16] For any value of f_{ion}, f has values that range from 1 when $f_{ion}=0$ to about 3 when $f_{ion}=1.0$. If we assume $f=2$, $\bar{\rho}=10^{-21}$ gm/cm^3, then the time for 50 percent equilibrium of neutron-decay electrons is 4×10^8 sec, or about 12 years. The equilibrium intensity may be obtained from $Nf\bar{\rho}/n=10^{-12}$ gm-sec/cm^3 (curve A, figure 3). If we assume the electron injection f and $\bar{\rho}$ as above, then $N=1.5\times10^{-4}$ electrons/cm^3. An approximate flux can be obtained by multiplying this figure by the most probable velocity of 2.2×10^{10} cm/sec, yielding a flux of 3.4×10^6 electrons/cm^2-sec for neutron-decay electrons above 30-kev energy as a representative value for the outer belt. It should be noted that electron-electron scattering does not scale as above since it is independent of f at these energies. However this effect is not as large as that of ionization energy loss or multiple scattering, and f is bounded between 1.0 and 3.0. Therefore the scaling error should not be large. The numerical problem of this paper was performed with $f=2.0$, and so it does not affect the example above. In figure 3, then, f should not be a variable but should be set equal to 2.0.

The possibility has been suggested that there are a large number of low-energy electrons in the outer radiation belt.[17] We have made a further calculation bearing on this low-energy group of electrons. A lifetime of about 1.5×10^7 sec is all hydrogen atmosphere of 1200 atoms/cm^3 average density, because of its range and multiple-scattering loss. Thus 20-kev electrons must be injected at a rate of 7×10^{-7} electrons/cm^3-sec to achieve equilibrium fluxes of 10^{11} electrons/cm^2-sec. This source strength is about 10^7 times that of the neutron β-decay source.

The spectrum calculated here is for neutron β-decay electrons trapped in the outer Van Allen radiation belt. We would expect, however, that the spectrum at lower altitudes would be different, even on magnetic field lines that extend into the upper belt. We attempted to see how this spectrum at lower altitude would differ from that in the upper altitudes by keeping track of the number and energy of those electrons that were Coulomb-scattered out of the upper radiation belt. These particles are the ones whose mirror altitude is lowered, and therefore they feed the spectrum at lower altitudes. Using these scattered particles as an input spectrum for lower altitudes and using a lifetime proportional to E^2 for electrons in the nitrogen-oxygen atmosphere of lower altitudes,[6] we obtain a flux spectrum as shown in figure 6. It should be emphasized that the geometry used is not that of the radiation belts. Essentially, we have put electrons in a box of hydrogen and calculated the equilibrium spectrum inside the box. We also calculated the energy spectrum of particles scattered out of this box and weighted them with the lifetime they would experience in the lower atmosphere. Kellog[18] has also calculated this leakage spectrum using an

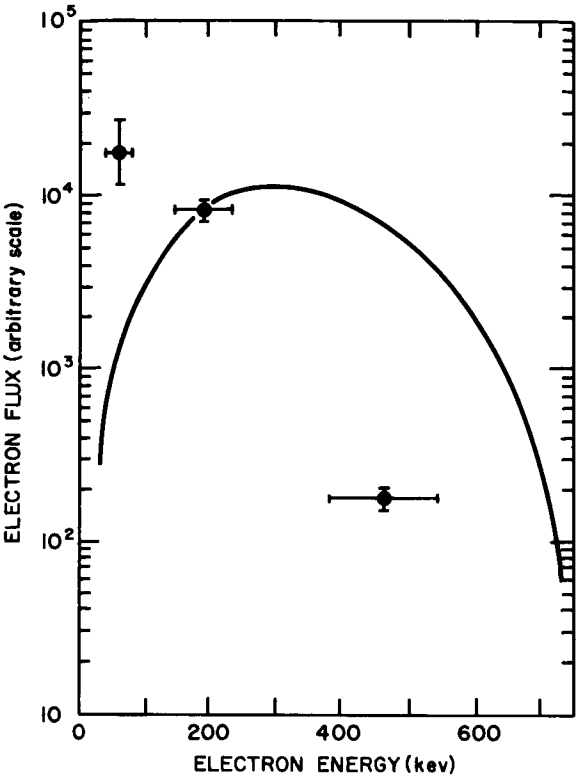


FIGURE 6.—Electron-leakage flux spectrum. This curve shows, as a function of energy, the sum of all particles that are lost from the upper Van Allen radiation belt by multiple-Coulomb scattering, weighed by their lifetime at lower altitudes and multiplied by their velocity to yield a flux. The curve has been drawn arbitrarily to pass through the second experimental point of Cladis, Chase, et al. (reference 21).

input spectrum of neutron β decay (our scattered spectrum has many more lower energy particles) and a loss mechanism of ionization energy loss [the E^2 lifetime of Welch and Whitaker[6] favors higher energy particles more strongly than ionization energy loss mechanisms]. As a result of these two balancing differences, about the same result is obtained.

Kellog argues that neutron decay is not an important source of electrons in the radiation belts. His argument, however, is based in large measure on electron flux measurements, which have, since then, been revised downward by as much as a factor of 10^3. O'Brien, Van Allen, et al.[19] have recently determined that there were $\approx 10^8$ electrons/cm^2-sec of energy greater than 40 kev in the outer belt on September 5, 1961. This was their highest reported flux; a time average of their values might be several times lower than this figure. The observed flux at these energies is about a factor of 10 larger than our calculated value. Since exospheric densities are quite uncertain and acceleration processes may substantially increase the lifetime of the neutron-decay electrons, it is not at all compelling that sources other than neutron decay are required.

Lenchek, Singer, and Wentworth[15] have recently calculated the equilibrium flux of outer-belt electrons to be expected from neutron β decay in a manner somewhat like ours here. They consider Coulomb scattering and ionization energy loss of the electrons. The shape of the energy spectra they calculate agrees quite well with ours, but they arrive at different conclusions about the origin of the electrons observed in the radiation belt. They conclude that less than 5 percent of the electrons seen by Holly and Johnson[20] in the inner radiation belt at $\approx$1100-km altitude are due to neutrons and that even a smaller fraction of the electrons seen by Cladis, Chase, et al.[21] at $\approx$1000 km in the outer belt are due to neutron decay. Lenchek and Singer state that neutron albedo contributes most if not all of the electrons above 400 kev but that few of the electrons of lower energy are made by neutrons.

We feel it is dangerous to base a general conclusion about the origin of all particles in the radiation belt on a few measurements at quite low altitudes. The analysis of these measurements depends on detailed information about the magnetic field of the earth's surface, the mirror-point altitude as a function of longitude, and the details of the atmosphere in a region where it is changing rapidly.

CONCLUSIONS

In this paper we have calculated the energy spectrum of electrons to be found in the outer belt if it were made of neutron β-decay electrons. Several experiments have given some information about the outer-belt spectrum, but the problem is not yet solved. The experiments involving bremsstrahlung will not be considered here because their interpretation is so difficult.

The Mechta experiment[22,23] indicated that there were about 1 percent as many electrons with

$E>500$ kev as with $E>50$ kev. Our calculated spectrum gives about 8 percent for the ratio of particles with $E>500$ to those with $E>50$ kev. Experimentally, there may be fewer electrons in the few-hundred-kev range than we calculated, but this is quite sensitive to the energy-threshold setting. If the Mechta experiment threshold had actually been 600 kev instead of 500 kev, we would agree with the 1 percent figure.

The leakage-electron energy spectrum measured by Cladis, Chase, et al.[21] is steeper and shows more low-energy electrons than can be explained by neutron decay. The fact that measurement was carried out at quite low altitudes may mean that processes or electron sources are involved here that do not contribute at high altitudes.

O'Brien, Van Allen, et al.[19] have recently reported on the results of their experiment on Explorer XII (1961 $\upsilon 1$). They have measured the electron spectrum in the middle of the outer belt by using several detectors sensitive in different energy ranges. The spectrum devised from these measurements for September 5, 1061, is shown on figure 7. Essentially, no electrons with $E>5$ Mev were seen, but some electrons in the range $1.6<E<5$ Mev were probably observed, although the interpretation of these experimental results is not unique. These high-energy electrons are very important for two reasons: First, they help explain results from several previous experiments that had seemed incompatible and, second, they cannot be made directly by neutron decay. There are not very many of these 1.6- to 5-Mev electrons; there were less than 1 percent of the total of September 5, and even less on other days.

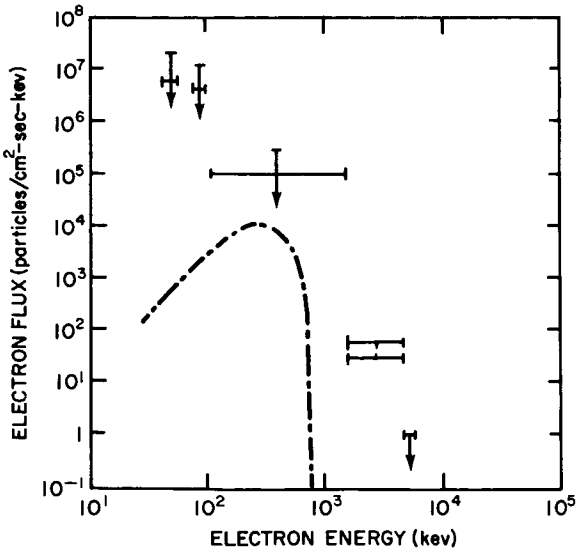


FIGURE 7.—Equilibrium flux of electrons attained by trapping 3×10^{-13} electron/cm³-sec in an atmosphere of hydrogen of average density 10^{-21} gm/cm³. The incident spectrum is assumed to be that of a neutron-decay electron; the hydrogen is assumed to be about 50% ionized. The experimental points and limits are those of O'Brien, Van Allen, et al. (reference 19) for September 5, 1961, when the electron flux was higher than average.

We must decide how similar the neutron β-decay electron energy spectrum is to the observed electron spectrum. Two features are to be compared: the shape, and the height—which amounts to comparing the flux. From figure 7, in comparing the *shape* of the two spectra (aside from the electrons above 1.6 Mev) it is not obvious whether the spectral shapes agree or not. There may be too many low-energy electrons observed to agree with the calculated spectrum. One thing is certain: There is not the very large low-energy peak that earlier measurements indicated.

In comparing the *height* of the spectra we probably should not take the experimental spectrum in figure 7 since it was the highest observed flux. We should decrease this by a factor of 3 or more. Using a representative time-average flux of $\approx3\times10^7$ electrons/cm²-sec for $E>40$ kev,[19] we find that our values are low by about one order of magnitude.

In this paper we have considered only neutrons made by galactic cosmic-ray protons. It is now quite certain that polar-cap protons contribute substantially to the low-energy trapped-proton spectrum[24,25,26] by generating neutrons in the polar atmosphere. Polar-cap protons will also, in the same way, add considerably to the trapped-electron flux. Lenchek[27] has recently estimated that 100 times as many neutrons are generated by polar-cap protons as by galactic cosmic-ray protons. If our calculated electron flux were increased by a factor of 100, the calculated flux would exceed the measured flux.

It appears from this that neutron decays produce a reasonable fraction of the outer-belt electrons, but other processes such as acceleration may be important. Electrons above 1 Mev may

be neutron β-decay electrons accelerated to these energies. We still need more definitive experiments to answer the question about whether any electron sources other than neutrons are needed to explain, in detail, the characteristics of the outer belt.

ACKNOWLEDGMENTS

We would like to thank Mrs. Kathryn Oliver, of the Lawrence Radiation Laboratory, for her assistance with the numerical calculations, and Professor Burton J. Moyer, University of California, for his support of this work. This work was done under the auspices of the U. S. Atomic Energy Commission.

REFERENCES

1. Freden, S. C., and White, R. S., "Particle Fluxes in the Inner Radiation Belt," *J. Geophys. Res.*, **65**(5): 1377–1383, May 1960.
2. Hess, W. N., "Van Allen Belt Protons from Cosmic-Ray Neutron Leakage," *Phys. Rev. Letters*, **3**(1): 11–13, July 1, 1959.
3. ———, "The Radiation Belt Produced by Neutrons Leaking Out of the Atmosphere of the Earth," *J. Geophys. Res.*, **65**(10): 3107–3115, October 1960.
4. ———, Canfield, E. H., and Lingenfelter, R. E., "Cosmic-Ray Neutron Demography," *J. Geophys. Res.*, **66**(3): 665–677, March 1961.
5. Christofilos, N. C., "Trapping and Lifetime of Charged Particles in the Geomagnetic Field," Univ. Calif., Radiation Lab. Rept. UCRL–5407, November 28, 1958.
6. Welch, J. A., Jr., and Whitaker, W. A., "Theory of Geomagnetically Trapped Electrons from an Artificial Source," *J. Geophys. Res.*, **64**(8): 909–922, August 1959.
7. Wentworth, R. C., MacDonald, W. M., and Singer, S. F., "Lifetimes of Trapped Radiation Belt Particles Determined by Coulomb Scattering," *Phys. Fluids*, 2(5): 499–509, September–October 1959.
8. Nelms, A. T., "Energy Loss and Range of Electrons and Positrons," Washington: U.S. Government Printing Office, 1956 (U.S. Nat. Bur. Standards Circular 577).
9. Bethe, H. A., and Ashkin, J., "Passage of Radiations Through Matter," in: *Experimental Nuclear Physics*, ed. by E. Segre, New York: John Wiley and Sons, 1953, Vol. 1, pp. 196–357.
10. "Symposium on Scientific Effects of Artificially Introduced Radiations at High Altitudes," *J. Geophys. Res.*, **64**(8): 865–938, August 1959; also *Proc. Nat. Acad. Sci.*, **45**(8): 1141–1228, August 15, 1959.
11. Møller, Chr., "Zur Theorie des Durchgangs Schneller Elektronen durch Materie," *Annalen der Physik*, **14**(5): 531–585, August 15, 1932; see also Jauch, J. M., and Rohrlich, F., "The Theory of Photons and Electrons," Reading, Mass.: Addison-Wesley, 1955.
12. Bethe, H., "Quantenmechanik der Ein- und Zwei-Elektronenprobleme," in: *Handbuch der Physik*, ed. by H. Geiger and K. Scheel, Berlin: Springer-Verlag, 1933, Vol. 24, Pt. 1, pp. 273–560.
13. Christofilos, N. C., "The Argus Experiment," *J. Geophys. Res.*, **64**(8): 869–875, August 1959.
14. Johnson, F. S., "The Exosphere and Upper F Region," *J. Geophys. Res.*, **65**(9): 2571–2575, September 1960.
15. Lenchek, A. M., Singer, S. F., and Wentworth, R. C., "Geomagnetically Trapped Electrons from Cosmic Ray Albedo Neutrons," *J. Geophys. Res.*, **66**(12): 4027–4046, December 1961.
16. Spitzer, L., Jr., "Physics of Fully Ionized Gases," New York: Interscience, 1956.
17. Arnoldy, R. L., Hoffman, R. A., and Winckler, J. R., "Observations of the Van Allen Radiation Regions During August and September 1959, Part 1," *J. Geophys. Res.*, **65**(5): 1361–1376, May 1960.
18. Kellogg, P. J., "Electrons of the Van Allen Radiation," *J. Geophys. Res.*, **65**(9): 2705–2713, September 1960.
19. O'Brien, B. J., and Van Allen, J. A., et al., "Absolute Electron Intensities in the Heart of the Earth's Outer Radiation Zone," *J. Geophys. Res.*, **67**(1): 397–403, January 1962; also State Univ. Iowa Rept. SUI 61–23, 1961.
20. Holly, F. E., and Johnson, R. G., "Measurement of Radiation in the Lower Van Allen Belt," *J. Geophys. Res.*, **65**(2): 771–772, February 1960.
21. Cladis, A. B., and Chase, L. J., Jr., et al., "Energy Spectrum and Angular Distributions of Electrons Trapped in the Geomagnetic Field," *J. Geophys. Res.*, **66**(8): 2297–2312, August 1961.
22. Vernov, S. N., and Chudakov, A. E., et al., "The Study of the Terrestrial Corpuscular Radiation and Cosmic Rays During the Flight of a Cosmic Rocket," *Soviet Phys.—Doklady*, **4**(2): 338–342, October 1959.
23. ———, and ———, "Radiation Measurements During the Flight of the Second Moon Rocket," *Soviet Phys.—Doklady*, **5**(1): 95–99, July–August, 1960.
24. Armstrong, A. H., Harrison, F. B., et al., "Charged Particles in the Inner Van Allen Radiation Belt," *J. Geophys. Res.*, **66**(2): 351–357, February 1961.
25. Naugle, J. E., and Kniffen, D. A., "Flux and Energy Spectra of the Protons in the Inner Van Allen Belt," *Phys. Rev. Letters*, **7**(1): 3–6, July 1, 1961.
26. ———, and ———, "The Flux and Energy Spectra of the Protons in the Inner Van Allen Belt," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961. II. *Joint Sessions*, Tokyo: Physical Society of Japan, 1962, pp. 118–122.
27. Lenchek, A. M., "On the Anomalous Component of Low-Energy Geomagnetically Trapped Protons," *J. Geophys. Res.*, **67**(6): 2145–2157, June 1962.

ELECTRONS FROM BOMB NEUTRON DECAY

JOHN KILLEEN

Lawrence Radiation Laboratory

WILMOT N. HESS

Goddard Space Flight Center

RICHARD E. LINGENFELTER

Institute of Geophysics and Planetary Physics, University of California

Calculations have been made of the initial electron distribution in space resulting from decay of neutrons emanating from a point source outside the earth's atmosphere. The albedo of neutrons from the earth's atmosphere and partial thermalization produces an important part of the electrons. About 80 per cent of the neutrons incident on the atmosphere are reflected. Since approximate normalization to the Starfish explosion shows that fluxes of less than 10^7 electrons cm^{-2} sec^{-1} are expected initially near the explosion site, this source does not contribute a large fraction of the observed artificial radiation belt.

INTRODUCTION

The purpose of this paper is to calculate the trapped-electron flux resulting from the decay of neutrons emanating from a point source. The best example of such a source is a nuclear explosion above the earth's atmosphere. Neutrons going out radially from the explosion decay by the reaction

$$n \rightarrow p + e + \bar{\nu}$$

with a mean life τ of 1000 seconds. The electrons resulting from the neutron decay have an energy spectrum nearly independent of the neutron velocity [Nakada, 1963].

Several recent high-altitude nuclear explosions have essentially been point sources of neutrons. It is of interest to evaluate the additions to the radiation belt resulting from neutron decay. It is known that some energetic particles were observed at large distances from the Starfish explosion of July 9, 1962. Ariel observed particles up to $L=6$ [Durney et al., 1962]. We will consider here sources both off and on the magnetic equator and study neutrons directly from the source and also albedo from the top of the atmosphere.

AN EQUATORIAL SOURCE

Starting with a neutron source on the magnetic equator the total flux F of neutrons passing through 1 cm^2 area at a point in space at a distance

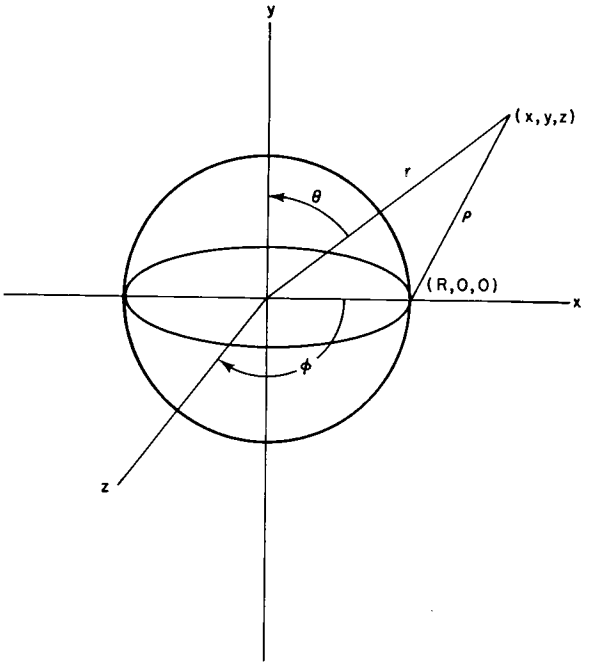


FIGURE 1.—Coordinate system.

ρ from the source is

$$F = M/4\pi\rho^2$$

where M is the total number of neutrons from the source. This neglects neutron albedo from the atmosphere, which will be considered later. The number of neutron decays per cubic centimeter at this point is

$$n_0 = M/4\pi\rho^2 v\tau \tag{1}$$

where v is the neutron velocity. Evaluating from figure 1,

$$\rho^2 = (x-R)^2+y^2+z^2 \tag{2}$$

where

$$\begin{aligned} x &= r\sin\theta\cos\varphi \\ y &= r\cos\theta\cos\varphi \\ z &= r\sin\varphi \end{aligned} \tag{3}$$

Substituting gives

$$\rho^2 = r^2+R^2-2rR\sin\theta\cos\varphi \tag{4}$$

Let $s=r/R$ and $\alpha=\rho/R$; then

$$n_0(s,\theta,\varphi) = \frac{M}{4\pi v\tau R^2\alpha^2} = \frac{M}{4\pi v\tau R^2}\left(\frac{1}{(1+s^2-2s\sin\theta\cos\varphi)^{1/2}}\right)^2 \tag{5}$$

The equation of a field line is

$$(\sin^2\theta_0)/R = (\sin^2\theta)/r \tag{6}$$

where θ_0 is the magnetic colatitude of the line at the earth's surface. Along a field line we have

$$s = (\sin^2\theta)/(\sin^2\theta_0) \tag{7}$$

The neutron decay density given in (5) is also the electron source density. Integrating equation 5 over longitude we get

$$\bar{n}_0(r,\theta) = \pi^{-1}\int_{-\varphi_0}^{+\varphi_0} n_0(r,\theta,\varphi)\,d\varphi = \frac{M}{2\pi^2 v\tau R^2}\int^{\varphi_0}\frac{d\varphi}{1+s^2-2s\sin\theta\cos\varphi}. \tag{8}$$

This represents the electron source strength after the electrons have had time to spread out in longitude. The value of φ_0 can be found from the fact that we consider only $x>R$, or $r\sin\theta\cos\varphi_0>R$; this gives

$$\varphi_0 = \arccos[1/(s\sin\theta)] \tag{9}$$

Equation 8 can be integrated directly to give

$$n_0(s,\theta) = \frac{M}{\pi^2 v\tau R^2}\left[(1+s^4+2s^2\cos 2\theta)^{-1/2}\arctan\left\{\frac{(1+s^4+2s^2\cos 2\theta)^{1/2}}{1+s^2-2s\sin\theta}\left(\frac{s\sin\theta-1}{s\sin\theta+1}\right)^{1/2}\right\}\right] \tag{10}$$

This has been evaluated to give Figure 2, using

$$k = M/\pi^2 v\tau R^2 = 1$$

This equation gives the density distribution of electrons produced along different field lines. Assuming that the electrons are emitted isotropically, we can calculate W, the distribution of mirror points of the electrons

$$W(\theta_0,\theta) = \frac{1}{2}\frac{dB}{dl}\frac{1}{B(\theta_0,\theta)}\int_{\pi/2}^{\theta}\frac{n_0(\theta_0,\theta')\,dl'}{[1-B(\theta_0,\theta')/B(\theta_0,\theta)]^{1/2}} \tag{11}$$

from which we can get the distribution of electron flux along a field line N by

$$N(\theta_0,\theta) = B(\theta_0,\theta)\int_{\theta}^{\theta_0}\frac{W(\theta_0,\theta_t)\,dl_t}{B(\theta_0,\theta_t)vT(\theta_0,\theta_t)[1-B(\theta_0,\theta)/B(\theta_0,\theta_t)]^{1/2}} \tag{12}$$

These expressions for W and N are developed in Hess and Killeen [1961]. They have been evaluated to give figures 3 and 4.

ALBEDO NEUTRONS

For a source of neutrons above the atmosphere we must consider not only neutrons coming upward from the source but also albedo from the top of the atmosphere. Neutrons initially directed downward into the atmosphere from the explosion site will suffer collisions in the upper atmosphere, and a majority of the neutrons will be scattered back out of the atmosphere. The intensity and energy spectrum of this albedo flux have been calculated from multigroup diffusion theory, described in detail by Lingenfelter et al. [1961].

The neutrons are assumed to come from an isotropic point source above the atmosphere, and separate calculations are made for neutrons with both a Maxwellian energy distribution at a temperature of 1 kev and a fission neutron energy spectrum. It is also assumed that the neutrons scatter isotropically at their first collision and that the density of first collisions decreases with altitude in the form $\exp(-x/L\mu)$, where x is the depth beneath the top of the atmosphere in grams per square centimeter, L is the collision mean free path (equal to about 4 cm^2/g for 1-kev neutrons and about 12 cm^2/g for fission neutrons), and μ is the cosine of the angle of incidence at the top of the atmosphere.

For normally incident neutrons, the albedo, or fraction of neutrons reflected, was calculated to be 0.74 for 1-kev neutrons and 0.63 for fission neutrons. Both values increase to unity with decreasing angle of incidence, as is shown in Figure 5. Integration of these two functions over μ gives a total albedo of 0.80 for neutrons with a Maxwellian distribution at 1 kev, leaving the explosion site in the downward hemisphere, and 0.73 for similar neutrons with a fission spectrum.

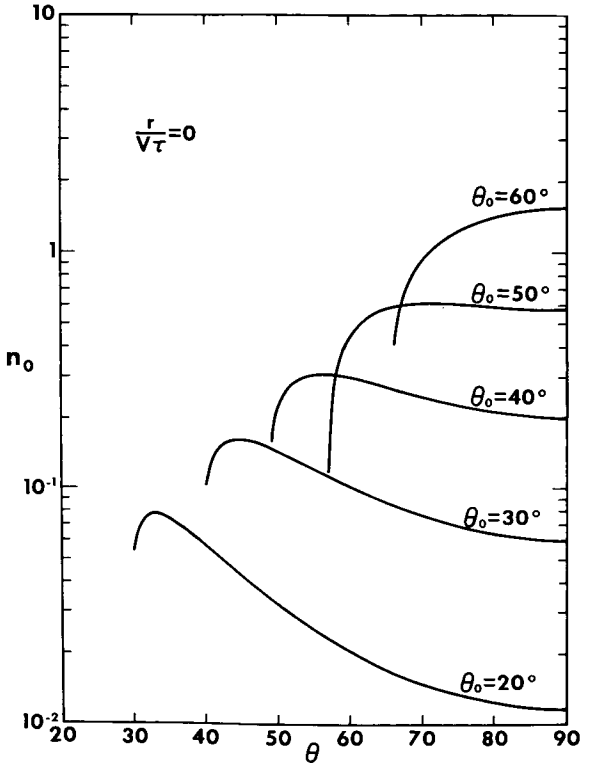


FIGURE 2.—Electron source density at various points in space for several field lines of colatitudes θ_0 resulting from decay of neutrons from a point source on the equator.

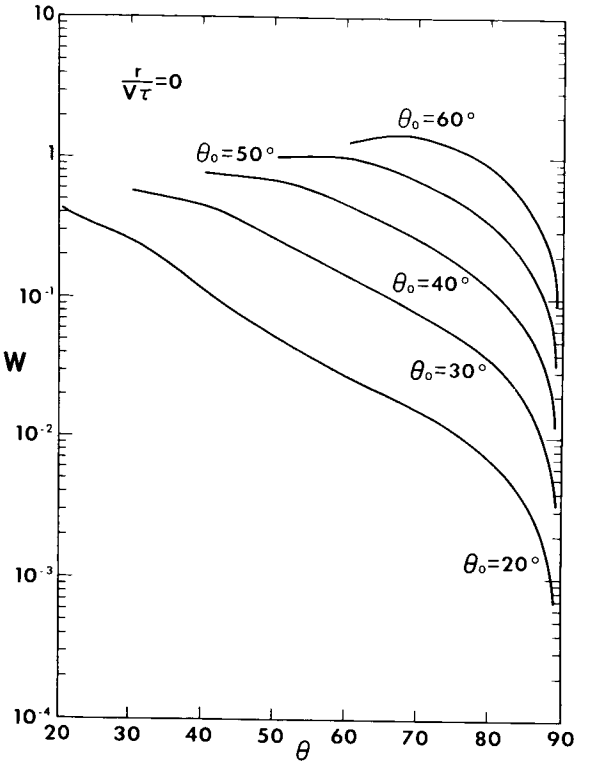


FIGURE 3.—Initial mirror point density of electrons at various points in space for several field lines of colatitude θ_0 resulting from decay of neutrons from a point source on the equator.

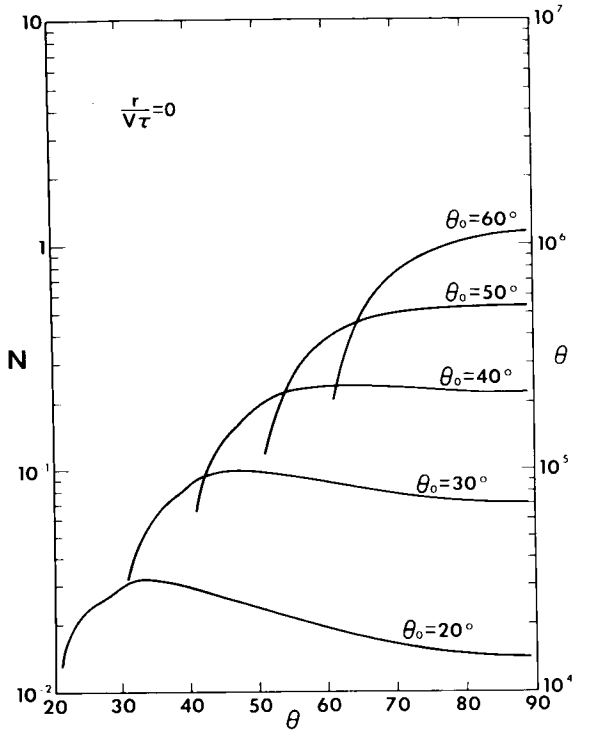


FIGURE 4.—Initial electron flux distribution along several field lines of colatitude θ_0 resulting from decay of neutrons from a point source on the equator for $k=M/\pi^2 v\tau R^2=1$.

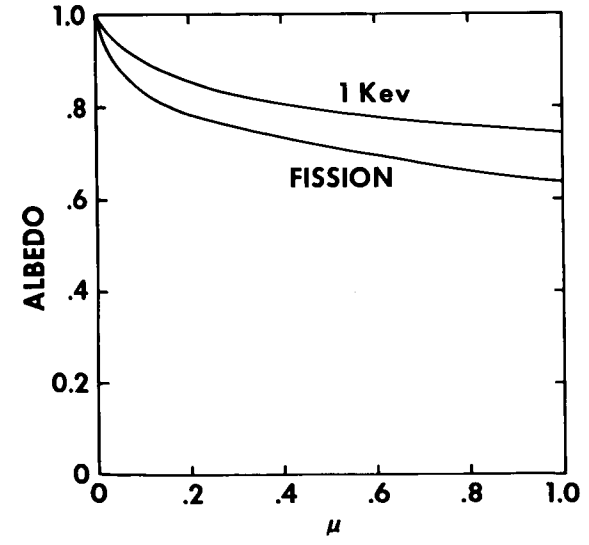


FIGURE 5.—The calculated albedo for neutrons from a point source above the atmosphere as a function of μ, the cosine of the angle of incidence at the top of the atmosphere, for neutrons with a Maxwellian energy distribution at 1 Kev and for a fission energy spectrum.

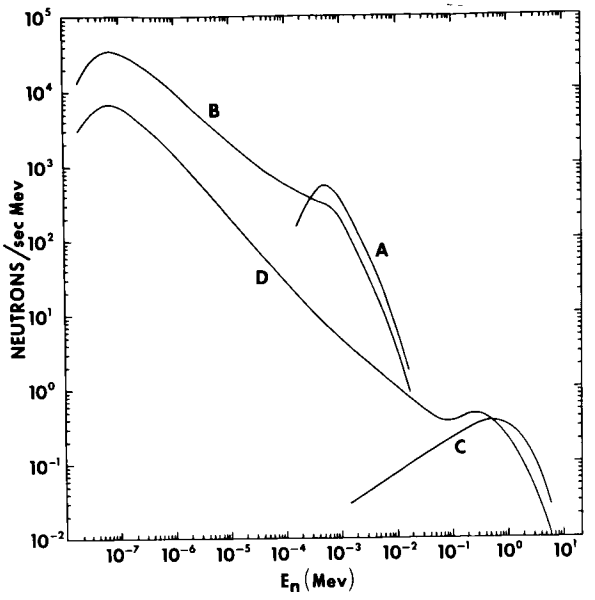


FIGURE 6.—Neutron energy spectrums for (a) a 1-Kev Maxwellian source, (b) the albedo from 1-Kev neutrons on the atmosphere, (c) a fission source, and (d) the albedo spectrum from fission electrons on the atmosphere.

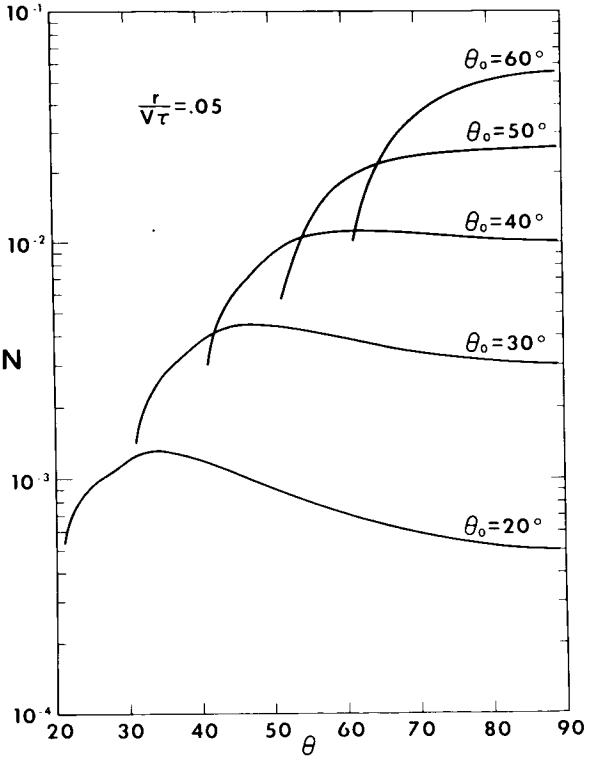


FIGURE 7.—Same as figure 4 except that decay of the neutron source is considered and $R/v\tau=0.05$ and $k=M/\pi^2R^8=1$.

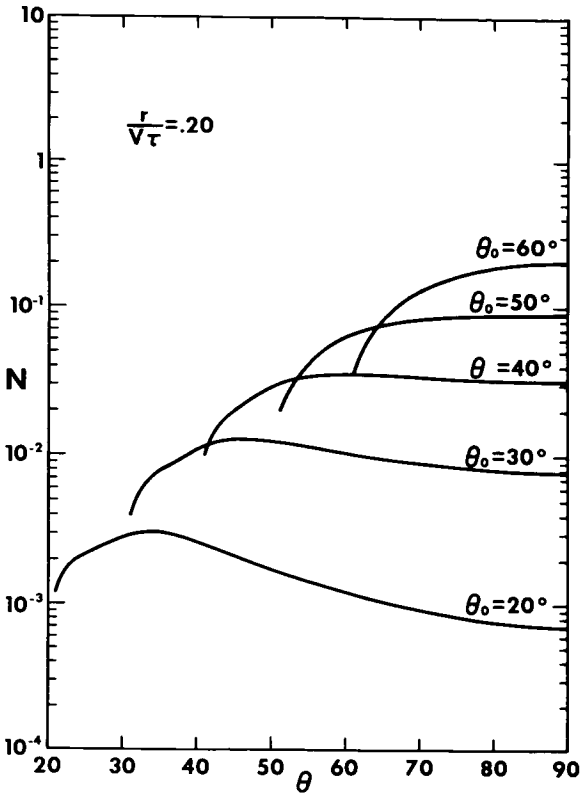


FIGURE 8.—Same as figure 7 except that $R/v\tau = 0.20$.

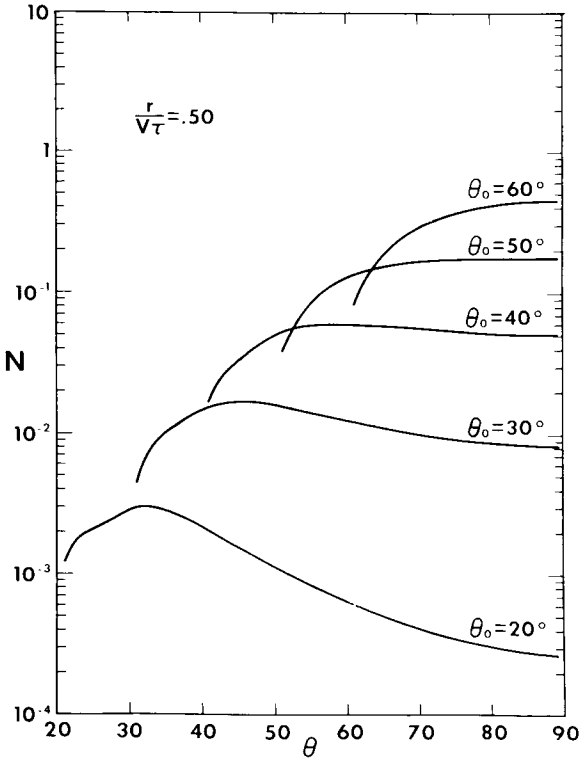


FIGURE 9.—Same as figure 7 except that $R/v\tau = 0.50$.

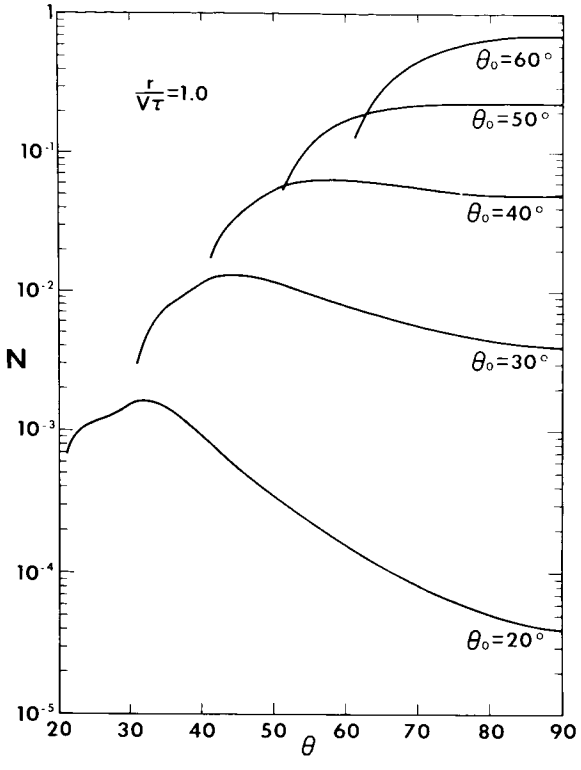


FIGURE 10.—Same as figure 7 except that $R/v\tau = 1.0$.

The calculated energy spectrums of the albedo fluxes from 1-kev and fission neutrons are shown in figure 6, where they are compared with the incident spectrums. Since diffusion theory is not rigorous within a couple of mean free paths of a boundary, the albedos thus calculated should probably have an uncertainty of about ±20 percent.

It is necessary to consider the neutron decay more completely in studying the albedo neutrons in the electron-volt energy range. In this case the neutron decay density at a distance ρ from

TABLE 1.—*Five Energy Groups*

$R/v\tau$	E	Energy range	f
0	∞	310 ev→∞	1.56
0.05	77 ev	31–310 ev	0.17
0.2	4.8 ev	3.1–31 ev	5.4×10^{-2}
0.5	0.8 ev	0.31–3.1 ev	1.6×10^{-2}
1.0	0.2 ev	0–0.31 ev	8.3×10^{-3}

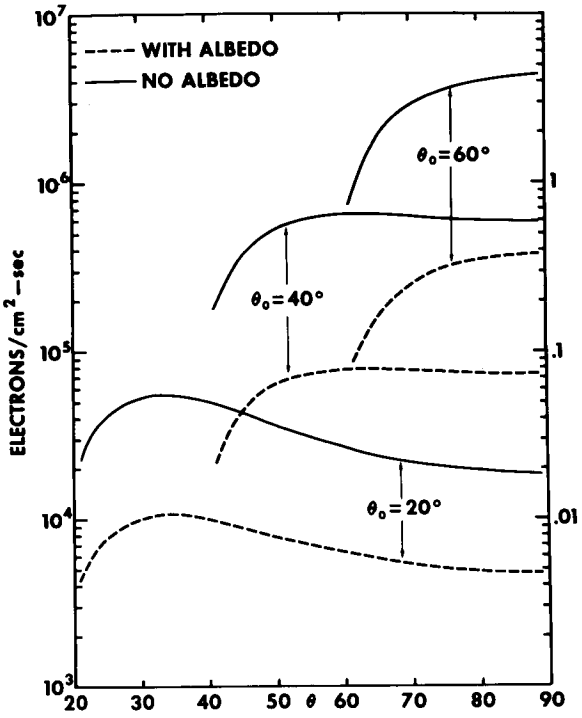


FIGURE 11.—Neutron decay electron fluxes expected from the Starfish explosion with and without albedo from the atmosphere.

the source is

$$n_0 = M/4\pi\rho^2 v\tau \exp(-\rho/v\tau) \qquad (13)$$

considering neutron decay between source and observer. The integration over φ is given by

$$\bar{n}_0(s,\theta) = \frac{M}{2\pi^2 v\tau R^2}\int_0^{\varphi_0} \frac{\exp(-R\alpha/v\tau)\, d\varphi}{1+s^2-2s\sin\theta\cos\varphi} \qquad (14)$$

This equation has been worked out for values of

$$\frac{R}{v\tau} = \frac{6.4\times10^8}{1.45\times10^6 (E_{ev})^{1/2}\times1000} = \frac{0.44}{(E_{ev})^{1/2}}$$

$$= 1,\ 0.5,\ 0.2,\ 0.05$$

The decay densities n have been transformed to electron flux distributions N in the same manner as before. Values of N are plotted for the four values of $R/v\tau$ in figures 7 to 10 for $M/\pi^2R^3 = 1$.

We can now apply our analysis to a specific case. The July 9 Starfish explosion, of about 1.4 MT, released of the order of 10^{26} neutrons. It was near enough the equator so that the current analysis is reasonable. Let us assume that 10 percent of these neutrons were thermalized in the bomb debris to a temperature of 1 kev, or $v = 0.45\times10^8$ cm/sec. It may not be a good assumption, but we will make it for lack of better information. We can evaluate the normalizing constant k by breaking the neutron energy spectrum in figure 6A and B down into five energy groups appropriate to the five values of $R/v\tau$ shown in Table 1. Then we evaluate five normalizing constants

$$k_i = (M/\pi^2R^2)\,(f_i/v_i\tau)$$

where f_i is the fraction of neutrons in the appropriate energy range given in the table (totaling 1.80 due to the albedo). Using these values of k to normalize the fluxes in figures 7 to 10, and then adding the results, gives the fluxes shown in figure 11. These data in figure 11 show the electron flux expected from Starfish with and without albedo neutrons. It has been calculated for 10 percent of the neutrons released appearing at 1 kev. This is only the injection flux, and the lower-altitude flux will decrease fairly rapidly as a result of coulomb scattering in the atmosphere. Calculations based on 100 per cent of the released neutrons having a fission spectrum give quite similar results because the higher total flux is offset by the smaller fraction decaying.

OFF-EQUATOR SOURCES

The Soviets have recently conducted high-altitude explosions in Siberia. To calculate the trapped-electron distribution resulting from neutron decay from such a source we can generalize the earlier expressions. Taking a source at colatitude η of coordinates ($R\sin\eta$, $R\cos\eta$, 0) and a point of observation at (x, y, z) as before, we can calculate the source-observer distance ρ by

$$\rho^2 = (x - R\sin\eta) + (y \pm R\cos\eta)^2 + z^2 \qquad (15)$$

The plus-or-minus sign here depends on whether the point of observation is in the northern (−) or the southern (+) hemisphere.

Using (15) gives a neutron decay density of

$$n_0 = \frac{M}{4\pi v\tau R^2}$$

$$\times (1+s^2-2s\sin\theta\cos\varphi\sin\eta \pm 2s\cos\theta\cos\varphi\cos\eta)^{-1} \qquad (16)$$

which is integrable as before to give

$$\bar{n}_0(s, \theta, \eta) = \frac{M}{4\pi^2 v\tau R^2} \int_{-\varphi_0}^{+\varphi_0} \frac{d\varphi}{a+b\cos\varphi} \quad (17)$$

The integral is broken into two parts, the northern hemisphere ($+\varphi_0$ to 0) and the southern hemisphere (0 to $-\varphi_0$), yielding

$$\int_0^{\varphi_0} \frac{d\varphi}{a+b\cos\varphi} = \frac{2}{(a^2-b^2)^{1/2}} \cdot \arctan\left[\frac{(a^2-b^2)^{1/2}}{a+b} \tan\frac{\varphi_0}{2}\right] \quad (18)$$

where $a=1+s^2$, $b=-2s\sin\theta\sin\eta \pm 2s\cos\theta\cos\eta$, and φ_0 can be evaluated from the fact that neutrons appear only on the side of the earth where the source exists. In the region where the source exists,

$$x > R\sin\eta + R(\cos^2\eta/\sin\eta) - y\cot\eta \quad (19)$$

Substituting for x and y gives the limit

$$\cos\varphi_0 = [s(\sin\theta\sin\eta \pm \cos\theta\cos\eta)]^{-1} \quad (20)$$

Here the plus sign is for the northern hemisphere.

REFERENCES

Durney, A. C., Elliot, H., Hynds, R. J., and Quenby, J. J., Satellite observations of the energetic particle flux produced by the high-altitude nuclear explosion of July 9, 1962, *Nature*, **195**, 1245–1248, 1962.

Hess, W. N., and Killeen, J., Spatial distribution of electrons from neutron decay in the outer radiation belt, *J. Geophys. Res.*, **66**, 3671–3680, 1961.

Lingenfelter, R. E., Canfield, E. H., and Hess, W. N., The lunar neutron flux, *J. Geophys. Res.*, **66**, 2665–2671, 1961.

Nakada, M. P., High-energy electrons in the radiation belt, *J. Geophys. Res.*, **68**, 47–53, 1963.

ON THE ANOMALOUS COMPONENT OF LOW-ENERGY GEOMAGNETICALLY TRAPPED PROTONS

A. M. LENCHEK

Goddard Space Flight Center

The energy spectrum of trapped protons shows a pronounced variation with latitude at energies below 30 Mev. An anomalous, low-energy component is found on relatively high magnetic shells in the inner zone. The interpretation of this anomalous component in terms of the decay of albedo neutrons generated in the polar caps by solar cosmic rays is examined. This interpretation requires that the inner shells, where a normal spectrum is seen, should be shielded from the polar source. The intensity of albedo neutrons as a function of energy, latitude, and zenith angle is computed. The injection rate into specific trapped orbits is then evaluated and it is shown that those orbits where no anomalous component exists cannot be populated by the solar cosmic ray mechanism while the orbits where the anomalous component is seen can be populated. When the assumption is made that there is adiabatic trapping in this low energy region, lifetimes are long compared to the average interval between large injection events. Therefore, the anomalous component may represent an equilibrium phenomenon rather than a transient. This component should extend into the outer belt.

LIST OF SYMBOLS

E	The energy of the secondary neutron and trapped proton.
$E_{max}(L)$	The maximum energy of a proton which remains trapped in the quiescent geomagnetic field.
$F(\omega)$	The equilibrium energy spectrum of differential intensity, given by equation 40 (dimensionless).
$I(\lambda', x)$	The omnidirectional integral flux of primaries (at latitude λ') which have energies above 10 Mev at depth x.
$I_0(\lambda')$	The integral over the upper hemisphere of the integral unidirectional intensity of primaries above 10 Mev.
$\bar{J}$	The integral unidirectional intensity of solar cosmic rays averaged over a solar cycle.
J_0	The integral unidirectional intensity of solar cosmic rays above 10 Mev, evaluated at a latitude at which the cutoff energy is less than 10 Mev.
$J(E)$	The integral unidirectional intensity of trapped protons with energies above E.
$J(>R, y)$	The directional integral intensity of protons with residual range greater than R.
$j(E, L, \alpha_0)$	The differential unidirectional intensity of trapped protons.
$j_n(E, \lambda', \phi)$	The differential unidirectional intensity of albedo neutrons.
$j_n{}^*(E)$	The vertical intensity of neutrons at the pole.
L	Magnetic shell parameter, equivalent to r_0/r_e in a dipole field.
L_{min}	The minimum shell accessible to protons injected by polar cap neutrons.
l	The attenuation length of neutrons passing through air.
$\mathfrak{M}$	The average mass per atom of the atmosphere.

$N(T)$	The probable number of neutrons emitted following an interaction with a proton energy T.
p_n/p_d	The neutron emission probability relative to deutron emission probability.
p_n/p_p	The neutron emission probability relative to proton emission probability.
p_n/p_α	The neutron emission probability relative to alpha particle emission probability.
$Q(E, \lambda', x)$	The production rate of neutrons of energy E at latitude λ' and depth x.
$q(E, L, \alpha_e; \lambda, \theta)$	The local injection rate giving number of protons injected/Mev-cm^3-ster-sec.
$\bar{q}(E, L, \alpha_0)$	The injection rate averaged over the spiral path.
$q(\lambda)$	The injection rate at λ averaged over one gyroperiod.
R	The range of solar proton in air.
R_0	$R(T_0)$.
R_1	$R(T_1)$.
$\bar{R}$	The mean range of the solar cosmic rays in model 1.
r	Geocentric distance.
r_e	The earth's equatorial radius.
r_0	The geocentric distance of a line of force measured in the equatorial plane.
$S(E)$	Evaporation spectrum.
$s(\alpha_0)r_0$	The length of the spiral path of a trapped particle, measured from the equatorial plane to a mirror point.
T	The energy of a solar proton.
T_0	The threshold energy of $N^{14}(p, n)O^{14}$ reaction.
T_1	The cutoff energy of solar cosmic rays assumed in model 2.
$T_{1/e}$	The $1/e$ lifetime, the time over which the energy decreases to $1/e$ times the initial value.
t_0	The duration of solar cosmic ray event.
t_n	The mean lifetime of the neutron.
$u = R_0/l \cos\phi$.	
v	Neutron velocity.
x	Atmospheric depth.
$y = x \sec\phi$	The slant depth of a point in the atmosphere.
α	The local pitch angle, the angle between the velocity vector and the magnetic field.
α_0	The equatorial pitch angle.
α-cone	The cone swept out on the globe by the gyration of the velocity vector.
δ-cone	A cone, of semiangle δ, drawn from a point in space to the earth; the cone is tangent to the earth's surface.
ϵ_3	The injection coefficient for orbits mirroring at NERV point number 3.
ϵ_4	The injection coefficient for orbits mirroring at NERV point number 4.
ϵ_{55}	The injection coefficient under the assumption that the solar cosmic rays reach a latitude of 55 degrees and higher.
ϵ_{60}	The injection coefficient under the assumption that the solar cosmic rays reach a latitude of 60 degrees and higher.
$\epsilon(L, \alpha_0)$	The injection coefficient, giving the rate of injection into an orbit in terms of the rate $j_n^*/t_n v$, corresponding to the decay rate of the vertical albedo intensity at the pole.
Θ	Nuclear temperature.
Θ'	The temperature appropriate to each stage of evaporation cascade.
θ	The phase angle between the velocity vector and the meridian plane.
λ	Magnetic latitude.
λ_C	The latitude of the boundary of the polar cap.

λ_m	Mirror point latitude.
λ'	The geomagnetic latitude of a point on globe from which an albedo neutron is emitted.
$\lambda^\dagger$	The highest latitude on earth's surface attained by the α-trace.
$\bar{\rho}$	The density of atoms-and-ions/cm^3 averaged over a trapped orbit.
$\bar{\rho}_3$	The average density for orbits mirroring at NERV point number 3.
$\bar{\rho}_4$	The average density for orbits mirroring at NERV point number 4.
σ_0	$\sigma(T)$ for $T>10$ Mev.
σ_t	The total cross section for neutron interaction in air.
$\sigma(T)$	The cross section for production of one neutron.
$\sigma_C(T)$	The cross section for compound nucleus formation as a function of proton energy.
$\Phi(\lambda', \phi)$	Defined by equation 5.
$\Phi_1(\lambda'\phi)$	$\Phi(\lambda', \phi)$ for model 1.
$\Phi_2(\lambda'\phi)$	$\Phi(\lambda', \phi)$ for model 2.
$\Phi_3(\lambda'\phi)$	$\Phi(\lambda', \phi)$ for model 3.
Φ^*	$\Phi(\lambda'>\lambda_C, \phi=0)$.
ϕ	The zenith angle of an albedo neutron upon leaving the atmosphere.
$\psi(x, \phi)$	The albedo angular distribution function, the ratio of albedo intensity at zenith angle ϕ to the vertical intensity at the pole.
$\Omega(L, \alpha_0, \lambda)$	Defined by equation 19.

INTRODUCTION

Observations of geomagnetically trapped protons have recently been extended to energies as low as 8 Mev[1] by use of the Nuclear Emulsion Recovery Vehicle (NERV). The energy spectrum shows a pronounced variation with position in space. The spectrum on magnetic shells crossing the geomagnetic equator at geocentric distances r_0 of approximately 1.5 earth radii r_e agrees well with theoretical predictions[2,3] based on injection by neutron albedo from galactic cosmic radiation.[4,5]

However, the spectrum observed on shells at $r_0>1.7$ r_e is anomalous in the sense that it is much steeper and more intense than the spectrum which results from "galactic" albedo, see figure 1. It has been suggested[1,6] that the neutron albedo generated in the polar caps by the arrival of low-energy solar cosmic rays may result in trapped protons. Some discussions of the characteristics of the protons resulting from this effect have already been given.[3,7] The object of this paper is to treat this mechanism more quantitatively than references 3 and 7 in an attempt to show that it can account for the NERV observations.

Three prominent features distinguish the solar cosmic rays from the galactic cosmic rays: (1) Solar cosmic rays arrive in intense bursts of short duration, the intensity typically decaying to normal in about a day, (2) arrive mostly at high latitudes, and (3) have a mean energy two or three orders of magnitude below that of the galactic radiation. Because of this last feature we may expect the number and energy spectrum of secondary neutrons to differ markedly from those produced by galactic cosmic rays: First, the high-energy spallation component should be almost totally absent. Second, the evaporation component should show a relatively lower temperature. Third, unlike the products of the global (i.e., galactic) radiation, most of the "polar" neutrons will be generated close to the top of the atmosphere (the range of a typical solar proton, with energy $T \lesssim 100$ Mev, is much shorter than the mean free path of the neutrons which have energies around 100 Mev).

Previous work on the albedo neutron energy spectrum[3,4,8,9,10,11] has been based on data referring to high-energy interactions involving primaries with energies in the Bev range. However, the primaries involved in the present paper have energies on the order of 10–100 Mev. We therefore must re-examine the neutron production process, concentrating on the reaction $N^{14}(p, n)O^{14}$.

In order to determine the spatial distribution

of the trapped protons produced by this mechanism it is necessary to examine in detail the efficiency of trapping as a function of r_0 and pitch angle. This, in turn, requires a knowledge of the angular distribution of the albedo intensity.

We therefore begin by examining the production of neutrons in the atmosphere by low energy cosmic ray protons. We then calculate the albedo intensity as a function of energy and zenith angle. Next we examine the injection coefficient as a function of energy, altitude, and pitch angle. It is found that the results of these calculations depend quite strongly on the form of the primary energy spectrum. However, the spectrum of the solar beam is variable from one event to another. Me therefore carry out the calculations for several different spectra.

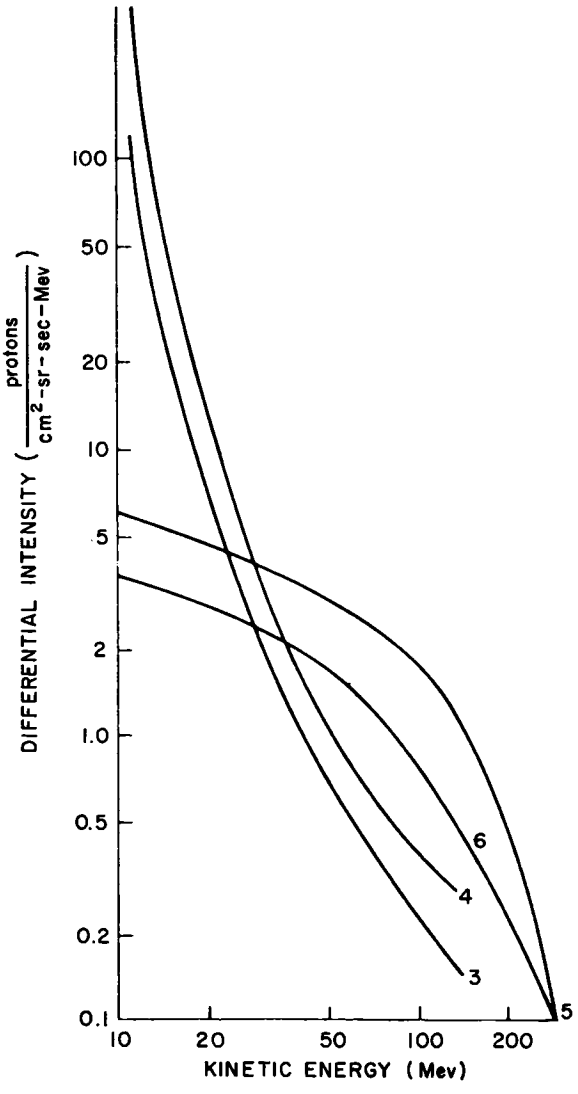


FIGURE 1.—Energy spectra of trapped protons on four different magnetic shells measured by the Nuclear Emulsion Recovery Vehicle program on September 19, 1960. Figure courtesy of Naugle and Kniffen. Altitudes and shell parameters L of the points are: point 3, alt. 1600 km, $L=1.79$; point 4, alt. $=1884$ km, $L=1.72$; point 5, alt. 1600 km, $L=1.54$; and point 6, alt. 1400 km, $L=1.47$.

PRODUCTION OF NEUTRONS

When protons with energies $T \lesssim 100$ Mev strike the atmosphere, the major source of secondary neutrons is the interaction of protons with N^{14}. Experimental data on the cross section for neutron production in this interaction is scanty. The neutron yield when 32 Mev protons bombard a thick nitrogen target is known.[12] The yield, 3.2×10^{-3} neutrons per proton, is comparable to the yield from oxygen. The energy spectrum of the neutrons produced in thin target experiments on nitrogen at 17.3 Mev[13] and at 31.5 Mev[14] shows several poorly resolved groups which tend to merge into a continuum at low energy. It is possible to represent the spectrum approximately in the form of an evaporation spectrum with a "temperature" around 4 Mev.

In this energy range, the mean free path of a nucleon is short compared to the nuclear diameter, suggesting that spallation may be less favored than compound nucleus formation. This observation, plus the evaporation-like nature of the neutron spectra, lead us to investigate the cross section for neutron production on the basis of evaporation theory. However, we rely on evaporation theory only to reveal the energy dependence and we use the observation of Tai et al.[12] for normalization.

Thus, we set $\sigma(T)=\sigma_C(T)N(T)$, where $\sigma_C(T)$ is the cross section for compound nucleus formation and $N(T)$ is the probable number of neutrons emitted. We compute $\sigma_C(T)$ in the semiclassical approximation,[15] which is valid for T greater than about twice the Coulomb barrier, that is, for $T>6$ Mev. We evaluate $N(T)$, using the theoretical results of LeCouteur,[16] by computing the relative probabilities P_n/P_p, P_n/P_α and P_n/P_d for emission of a neutron or proton;

neutron or alpha particle; and neutron or deuteron. Multiple particle emission is considered up to three-particle cascades although, since all the required separation energies are not available, triple cascades can only be estimated. In computing cascades it is assumed that all particles except the last are emitted with the average energy of those emitted in each stage. The relation between the nuclear temperature Θ and the excitation energy is assumed to be given by the usual degenerate Fermi-gas model of the excited nucleus.

It is found that proton emission is most likely in the first stage of de-excitation. The calculated $\sigma(T)$ rises rapidly as T is increased above the threshold $T_0=6.4$ Mev, and then levels off at about 10 Mev, attaining a maximum value of only about 70 mb. Multiple emissions tend to increase this figure at higher energies. The data of Tai et al. can be fitted by assuming that $\sigma(T)=0$ for $T<10$ Mev and $\sigma(T)=\sigma_0=80$ mb for $T>10$ Mev. We shall therefore adopt this simple model for $\sigma(T)$.

We assume the evaporation spectrum is given by

$$S(E)=E\Theta^{-2}\exp(-E/\Theta)\ \text{neutrons/Mev} \quad (1)$$

per interaction, where E is the neutron energy. In cascades in which more than one neutron is emitted, Θ is an average value found by assuming that all particles except the last are emitted with $E=2\Theta'$ (Θ' is the temperature appropriate to each state), weighting the several Θ' values by the fraction of the neutrons belonging to each stage, and summing over stages. It is then found that Θ is a quite slowly varying function of T, varying only between 3 and 5 Mev as T varies from threshold to 50 Mev. We may therefore take $\Theta=4$ Mev as an average, independent of T. This is an important simplification since it implies that the neutrons are generated with the same spectrum at all atmospheric depths.

We therefore conclude that (1) the neutron production is adequately represented (approximately) by a constant cross section of 80 mb with an effective threshold of 10 Mev, (2) the energy spectrum of the neutrons is Maxwellian with temperature 4 Mev, and (3) the production angular distribution is isotropic.

ALBEDO INTENSITY

We turn now to the question of the differential unidirectional intensity of neutrons leaving the atmosphere. We must compute the production rate as a function of atmospheric depth and, assuming for simplicity an exponential attenuation of the emerging beam, compute the escaping intensity as a function of energy and zenith angle.

The production rate at atmospheric depth x and at geomagnetic latitude λ' is

$$Q(E,\lambda',x)=\frac{\sigma_0 I(\lambda',x)\ S(E)}{\mathfrak{M}}\ \frac{\text{neutrons}}{\text{Mev-gm-sec}}, \quad (2)$$

where $\mathfrak{M}$ is the average mass per atom of the atmosphere and $I(\lambda', x)$ is the omnidirectional integral flux of primaries which have energies above 10 Mev at depth x.

The differential unidirectional intensity of albedo neutrons leaving the atmosphere at the zenith angle ϕ is obtained by integrating $Q(E, \lambda', x)$ over a slant path, assuming a plane stratified atmosphere:

$$j_n(E,\lambda',\phi)=\frac{1}{4\pi\cos\phi}\int_0^\infty Q(E,\lambda',x)\exp\left[\frac{-x}{l\cos\phi}\right]dx\ \frac{\text{neutrons}}{\text{Mev-cm}^2\text{-ster-sec}}. \quad (3)$$

Here l is the attenuation length, which we approximate by $\mathfrak{M}/\sigma_t$ where σ_t is the total cross section in air.[17] In the energy range in which we are interested (about 5–30 Mev), we can assume $l=15$ gm/cm^2, independent of energy. Then

$$j_n(E,\lambda',\phi)=\frac{\sigma_0 l}{4\pi\mathfrak{M}}\ I_0(\lambda')\ S(E)\ \Phi(\lambda',\phi), \quad (4)$$

where

$$\Phi(\lambda',\phi)=\frac{1}{l\cos\phi}\int_0^\infty\left[\frac{I(\lambda',x)}{I_0(\lambda')}\right]\exp\left(\frac{-x}{l\cos\phi}\right)dx. \quad (5)$$

Here $I_0(\lambda')$ is the integral over the upper hemisphere of the integral directional intensity of primaries above 10 Mev.

Let $j_n^*(E)$ denote the vertical intensity of neutrons at the pole. The boundary of the polar cap is at latitude λ_C, the latitude at which the geomagnetic cutoff is 10 Mev. We shall express $j_n(E, \lambda', \phi)$ in terms of $j_n^*(E)$ as

$$j_n(E,\lambda',\phi) = j_n^*(E)\,\Psi(\lambda',\phi)\,. \qquad (6)$$

Thus,

$$\Psi(\lambda',\phi) \equiv \frac{\Phi(\lambda',\phi)}{\Phi(\lambda' > \lambda_C,\ \phi = 0)}\,.$$

We shall denote $\Phi(\lambda'>\lambda_C,\ \phi=0)$ by Φ^*. Then

$$j_n^*(E) = \frac{\sigma_0 J_0 \ell\, S(E)\, \Phi^*}{2\pi}\,,$$

where J_0 is the integral directional intensity (in particles/cm²-ster-sec) of solar cosmic rays above 10 Mev, evaluated at a latitude where the cutoff energy is less than 10 Mev. We assume J_0 is isotropic; hence $J_0=I_0/2\pi$. Before we can evaluate the albedo distribution function $\Psi(\lambda', \phi)$ we must discuss the spectrum of the solar beam, since this determines $I(\lambda', x)/I_0(\lambda')$ appearing in equation 5.

THE SOLAR BEAM

Until recently no direct observations of the spectrum of solar cosmic rays were available. However, the spectra during the events of September 3, 1960, and November 12–15, 1960, were directly measured.[18,19] In the September 3 event, a Geiger counter and nuclear emulsions carried on the same rocket yielded an integral proton spectrum that was almost flat up to about 100 Mev with an intensity $J_0\sim19$ particles/cm-ster-sec above 22 Mev. In the November event the integral spectrum was close to $T^{-1.7}$ (from 10–100 Mev) with an intensity $J_0 \gtrsim 2\times10^3$ particles/cm²-ster-sec above 10 Mev. In both cases the intensity was isotropic over the upper hemisphere.

EVALUATION OF THE DISTRIBUTION FUNCTION

We shall consider three cases: (1) exponential variation of Q with depth, (2) primary spectrum flat up to 100 Mev, and (3) primary spectrum $J_0 \propto T^{-1.7}$.

Model 1

We assume

$$\frac{I(\lambda',x)}{I_0(\lambda')} = e^{-x/\bar{R}}\,, \qquad (8)$$

where $\bar{R}$ is the mean range of the solar cosmic rays. Inserting into equation 5, we obtain

$$\Phi_1(\lambda',\phi) = \frac{\bar{R}}{\ell}\left(\cos\phi + \frac{\bar{R}}{\ell}\right)^{-1}\,. \qquad (9)$$

Taking $\bar{R}$ to be the range of a representative primary ($T<100$ Mev), we find that $\bar{R}$ will be small compared with l, giving an anisotropic albedo with a maximum at the horizontal.

Model 2

Assume an isotropic intensity $J_0=$constant up to T_1 and $J_0=0$ above T_1. Convert to a range spectrum and transform to a depth x by noting that at the slant depth $y=x\sec\phi$ the directional integral intensity $J(>R, y)$ of protons with residual range greater than R equals the primary intensity $J_0(R+y)$ for $R+y<R(T_1)$ and is zero for $R+y>R(T_1)$. Denote $R(T_1)$ by R_1. Then integrate over the hemisphere to obtain

$$\frac{I(\lambda',x)}{I_0(\lambda')} = 1 - \frac{x}{R_1 - R_0} \quad \text{for} \quad x \le R_1 - R_0 \equiv x_0 \qquad (10)$$

and

$$I(\lambda', x) = 0 \quad \text{for} \quad x > x_0.$$

Here $R_0=R(T_0)=0.15$ gm/cm² and $R_1=R(T_1)=7.5$ gm/cm² assuming $T_1=100$ Mev.

Inserting equation 10 into equation 5, we easily find

$$\Phi_2(\lambda',\phi) = 1 - \frac{\ell\cos\phi}{x_0}\left(1 - e^{-\frac{x_0}{\ell\cos\phi}}\right). \qquad (11)$$

Since $x_0\sim l/2$ we again find a moderate anisotropy with $\Phi_2(\lambda', \pi/2)/\Phi_2(\lambda', 0)=4.71$.

Model 3

Assume $J_0=AT^{-1.7}$. Using the familiar relation $R(T)=\text{constant}\times T^{1.7}$ we then find

$$\frac{I(\lambda',x)}{I_0(\lambda')} = 1 - \frac{x}{R_0}\ln\frac{R_0 + x}{x} \qquad (12)$$

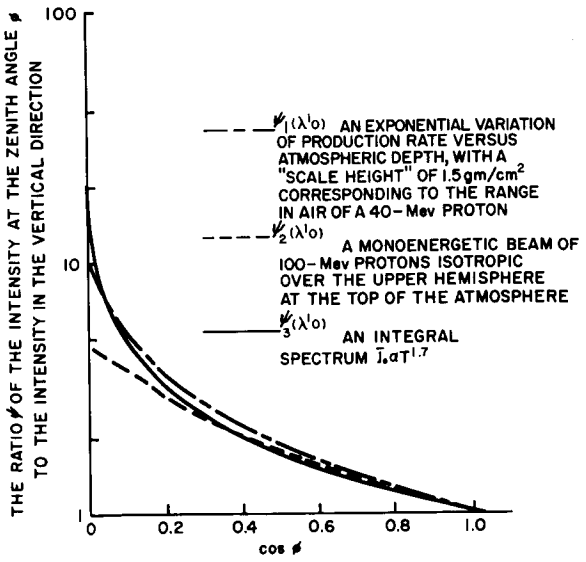


FIGURE 2.—Angular distribution functions of albedo neutrons generated by low-energy cosmic rays. In each case a constant cross section with threshold 10 Mev is used. Exponential attenuation of the emerging beam, with attenuation length 15 gm/cm², and a plane atmosphere are assumed.

and

$$\Phi_3(\lambda',\phi) = 1 + e^{u}\mathrm{Ei}(-u)\left[\frac{1}{u} - 1\right] - \frac{1}{u}(\gamma + \ln u), \qquad (13)$$

where $u = R_0/l \cos\phi$, $-Ei(-u)$ is the logarithmic integral and $\gamma = 0.577$ is Euler's constant. We find $\Phi_3(\lambda', 0) = 0.025$ and $\Phi_3(\lambda', \pi/2) \approx 1$.

Thus, in each case we find $\Phi(\lambda', 0) \sim R/l$, where R is the range of an average primary, and $\Phi(\lambda', \pi/2) \approx 1$.

The functions $\Psi(\lambda', \phi)$ obtained from the above models are shown in figure 2. In each case it is assumed that $\Psi(\lambda', \phi)$ vanishes for $\lambda' < \lambda_C$ and is independent of λ' for $\lambda' > \lambda_C$. This is a good approximation, since the latitude interval over which the cutoff energy varies from 10 to 100 Mev is quite narrow (the Störmer cutoff is 10 Mev at $\lambda' = 72$ degrees and is 100 Mev at $\lambda' = 66$ degrees, assuming vertical incidence) and the contribution from primaries with energies above 100 Mev is assumed to be negligible. However, cutoff rigidities are often reduced during cosmic ray increases. Cosmic rays with energies as low as 120 Mev have, at times of magnetic activity, been observed at Minneapolis.[20] The geomagnetic latitude of Minneapolis, defined so that the vertical cutoff is 14.9 $\cos^4\lambda' Bv$, is 57.4 degrees[21] at quiet times although protons of this energy should normally be excluded below about 65 degrees. We therefore carry out the calculation for $\lambda_C = 60$ and 55 degrees.

INJECTION COEFFICIENTS

The basic principle determining the rate of injection into a given orbit is that injection occurs only over those segments of the orbit in which the orbit points toward the source.[9] The decay proton maintains the direction of flight of the parent neutron. As the orbit spirals through space, the velocity vector of a particle in that orbit sweeps out a cone called the α-cone, where α is the local pitch angle. Over some parts of the orbit, the α-cone may miss the earth entirely. The intersection of the α-cone with the earth is called the α-trace. (Only a fraction of the α-trace may fall within the polar cap.) The albedo intensity varies from point to point along the α-trace depending on the zenith angle ϕ made by the velocity vector, and the magnetic latitude λ' of the point of intersection with the earth. Thus, ϕ and λ' may be regarded as "α-trace coordinates."

Since particles are trapped for times that are long compared to bounce periods, it is sufficient to calculate the average injection rate over the orbit. We label an orbit by its coordinates in the equivalent dipole coordinate system discussed by McIlwain.[22] Thus, an orbit is determined by $L = r_0/r_e$, where r_e is the earth's radius, and by the equatorial pitch angle α_0.

At a given point along the orbit, the position of the velocity vector relative to the meridian plane is denoted by θ, the phase angle. The local injection rate at latitude λ, energy E, and phase angle θ is then

$$q(E, L, \alpha_0; \lambda, \theta) = \frac{j_n(E, \lambda', \phi)}{t_n v}, \qquad (14)$$

where λ' and ϕ are functions of L, α_0, λ and θ. In equation 14, t_n is the mean lifetime of the neutron (including relativistic time dilation) and

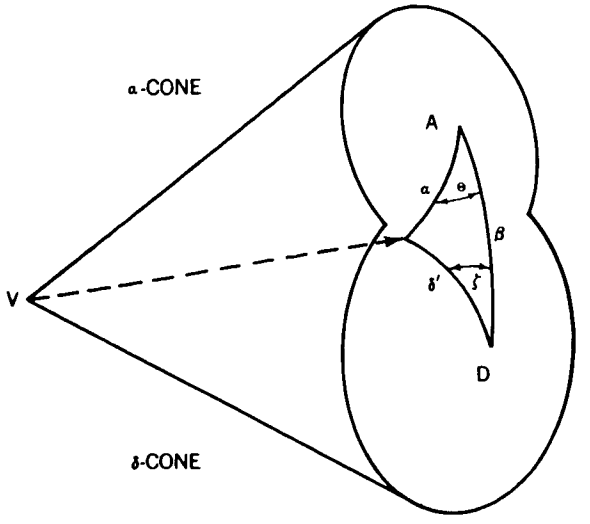


FIGURE 3a.—Two cones are constructed, each with its vertex at V, the guiding center. The α-cone is swept out by the gyrating velocity vector. The δ-cone is the cone subtended by the earth. The point A marks the axis of the α-cone; D marks the axis of the δ-cone. The angle between the two axes is β. When the two cones overlap, the end of the velocity vector, shown as a dashed line, traces out an arc within the δ-cone which we call the α-trace (not shown). The arc AD is a segment of the meridian. The angle θ is the phase angle; as θ varies, the angles δ' and ζ also vary but α and β remain fixed.

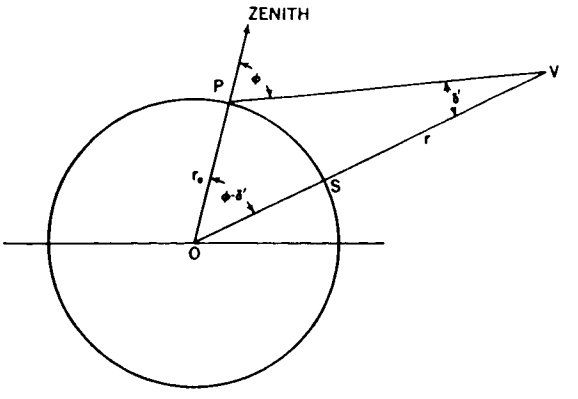


FIGURE 3b.—The circle represents the earth. The line VP is the velocity vector. The point S is the intersection of the axis of the δ-cone with the earth's surface.

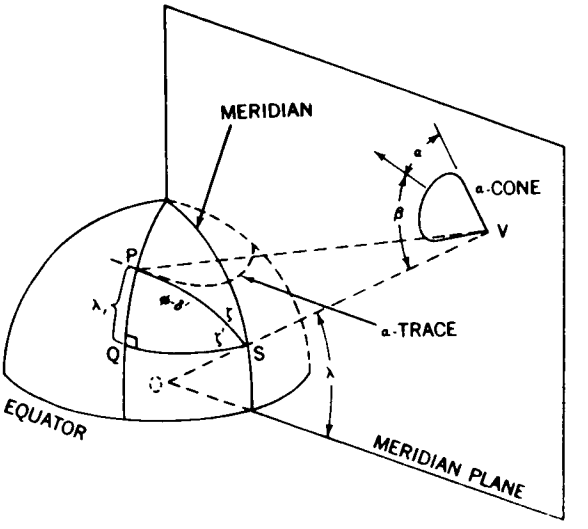


FIGURE 3c.—The axis of the δ-cone passes through the point S. The arc SQ is drawn along a line of constant latitude. The triangle of figure 3b lies in a plane containing the points P, S, and O, the center of the earth.

v is the neutron's velocity. Assuming that a negligible number of neutrons decay enroute, no correction is necessary for the distance from the earth. That is, since j_n is the differential unidirectional intensity, its value at the point of decay equals its value at the earth.

The quantity q must now be integrated over the spiral path. (The length of the spiral path may be expressed in terms of the distance r_0 as $s(\alpha_0)r_0$ where $s(\alpha_0)$ is dimensionless.)

Since the gyrofrequency is large compared to the bounce frequency $v/4s(\alpha_0)r_0$, we may average over the phase angle (i.e., over a gyroperiod) before integrating over $s(\alpha_0)$. Thus, the average injection rate $\bar{q}(E, L, \alpha_0)$ is

$$\bar{q}(E,L,\alpha_0) = \frac{1}{s(\alpha_0)} \int_0^{\lambda_m} \langle q(E,L,\lambda) \rangle \frac{ds}{d\lambda} \, d\lambda , \quad (15)$$

where

$$\langle q(E,L,\lambda) \rangle = \frac{1}{2\pi} \int_0^{2\pi} q(E,L,\lambda,\theta) \, d\theta \quad (16)$$

is the average over a gyroperiod, holding λ fixed. We denote the mirror point latitude by λ_m.

We now use equation 6 to express $\bar{q}$ in terms of j_n^*:

$$\bar{q}(E,L,\alpha_0) = \frac{\epsilon(L,\alpha_0) \, j_n^*(E)}{t_n v} \quad (17)$$

where the injection coefficient $\epsilon(L, \alpha_0)$ is

$$\epsilon(L,\alpha_o) = \frac{1}{s(\alpha_o)} \int_o^{\lambda_m} \Omega(L,\alpha_o,\lambda) \frac{ds}{d\lambda} d\lambda \quad (18)$$

and

$$\Omega(L,\alpha_o,\lambda) = \frac{1}{\pi} \int_o^{\pi} \Psi(L,\alpha_o,\lambda,\theta)\, d\theta . \quad (19)$$

where $\Psi(L, \alpha_0, \lambda, \theta)$ is the albedo angular distribution function defined by equation 6, now expressed in orbit coordinates L, α_0, λ, and θ.

Our task now is to relate the orbit coordinates L, α_0, λ, and θ to the α-trace coordinates ϕ and λ'.

For a given orbit, defined by L and α_0, and for some point along the line of force, defined by the latitude λ, we may construct the α-cone and examine its intersection with the earth (see figure 3). If λ is held fixed and the phase angle θ is varied, the velocity vector will trace out the α-trace on the earth's surface. The latitude λ' at which the vector intersects the earth is given by

$$\lambda' = \lambda + \lambda_1, \quad (20)$$

where $\lambda_1 = \lambda_1(\theta, \lambda, L, \alpha_0)$. The zenith angle ϕ at this point is related to the orbit coordinates by

$$\sin\phi = L \cos^2\lambda \sin\delta', \quad (21)$$

where

$$\cos\delta' = \cos\alpha \cos\beta + \sin\alpha \sin\beta \cos\theta, \quad (22)$$

$$\sin\alpha = \frac{(4 - 3\cos^2\lambda)^{1/4} \sin\alpha}{\cos^3\lambda}, \quad (23)$$

and

$$\tan\beta = \frac{1}{2}\cot\lambda . \quad (24)$$

Having determined ϕ and δ' from equations 21 through 24, we may then find λ_1 from

$$\sin\lambda_1 = \cos\zeta \sin(\phi - \delta') , \quad (25)$$

where

$$\cos\zeta = \frac{\cos\alpha - \cos\delta' \cos\beta}{\sin\delta' \sin\beta} . \quad (26)$$

With these relations we are able to compute ϕ and λ' everywhere along the α-trace; thus, Ω can now be evaluated from equation 19. The integration of equation 18 is then carried out with the aid of the relation

$$\frac{ds}{d\lambda} = (4 - 3\cos^2\lambda)^{1/2} \sec\alpha \cos\lambda . \quad (27)$$

The mirror point latitude is related to α_e by

$$\sin\alpha_e = \frac{\cos^3\lambda_m}{(4 - 3\cos^2\lambda_m)^{1/4}} , \quad (28)$$

and a close approximation to $s(\alpha_0)$ is given by[23]

$$s(\alpha_o) \approx 1.38 - 0.32 \left(\sin\alpha_o + \sin^{1/2}\alpha_o\right) . \quad (29)$$

When the α-cone does not intersect the earth, Ω is zero. For an orbit such that $\Omega(L, \alpha_0, \lambda) = 0$ for $\lambda = 0$, the latitude λ_2 such that the α-cone first touches the earth may be obtained from the implicit relation

$$\alpha(\lambda_2) + \delta(\lambda_2) = \beta(\lambda_2) , \quad (30)$$

where

$$\sin\delta = \frac{\sec^2\lambda}{L} . \quad (31)$$

The angle δ is the semiangle of the cone drawn from the orbit point to the earth, the cone being tangent to the earth's surface.

In general, only a part of the α-cone intersects the earth. In this case, the upper limit on the integral in equation 19 is replaced by θ', where

$$\cos\theta' = \frac{\cos\delta - \cos\alpha \cos\beta}{\sin\alpha \sin\beta} . \quad (32)$$

The angle θ' is related to the injection coefficient $\bar{\eta}_0$ calculated by Singer[9] through

$$\bar{\eta}_o = \frac{1}{s(\alpha_o)} \int_o^{\lambda_m} \eta(\lambda) \frac{ds}{d\lambda} d\lambda \quad (33)$$

and

$$\eta(\lambda) = \frac{\theta'}{\pi} . \quad (34)$$

If we assume an isotropic albedo and intensity uniform with respect to latitude, then ϵ reduces to $\bar{\eta}_0$.

The injection coefficients have been evaluated for the orbits which mirror at the points in space where the NERV observations were made. The

TABLE 1.—*Injection Coefficients for Orbits Mirroring at the Locations of the NERV Observations*

L=equivalent geocentric distance in equatorial plane; α_0=equatorial pitch angle; λ_2=latitude of particle when the α-cone first intersects the earth; $\lambda^\dagger$= highest latitude on earth's surface attained by the α-trace; ϵ_{60}=injection coefficient assuming solar cosmic rays reach latitude 60 degrees; ϵ_{55}=injection coefficient assuming solar cosmic rays reach latitude 55 degrees.

Point Number	L	cos α_0	λ_2	$\lambda^\dagger$	ϵ_{60}	ϵ_{55}
3	1.79	0.866	11.0	66.1	0.15	0.29
4	1.72	0.832	9.0	62.3	0.10	0.18
5	1.54	0.775	5.5	54.5	0	0
6	1.47	0.755	2.9	48.3	0	0

coordinates of these points are given in Table 1 along with the coefficients calculated from the albedo distribution of Model 3. The most important feature of the result is that the injection coefficient is zero at points 5 and 6. These are the points at which the anomalous component is absent. It can be shown that for a given value of λ_C there is a minimum value of L for which any injection is possible (for any value of α_0). In Table 2 we give $L_{\min}$ as a function of λ_C. Note that for $\lambda_C \approx 60$ degrees, the boundary of the accessible region is at $L=1.65$; this is consistent with the observations.

The vanishing of the injection coefficient at the two inner points argues against an interpretation of the anomalous component in terms of either direct injection from the solar plasma cloud or local acceleration. Both of these mechanisms seem unlikely to produce trapped radiation with such a sharp lower boundary. The polar albedo mechanism, however, leads to a sharp lower boundary in just the correct range of L.

TABLE 2.—*Minimum Injection Distance $L_{\min}$ as a Function of the Latitudes of the Boundary of the Polar Cap λ_C (Assuming Infinite Atmospheric Density Below 1270 km)*

λ_C (degrees)	$L_{\min}$ (earth radii)
0	1.20
10	1.22
20	1.27
30	1.32
40	1.38
50	1.46
60	1.65
70	1.88
80	2.30

The reason for the absence of injection at the inner points may be more clearly seen by considering the behavior of the α-cone or the α-trace as λ varies. All four orbits are of the class in which the α-cone does not intersect the earth when the vertex of the cone is at the equatorial plane. As λ increases, the α-cone tilts toward the earth and grows wider. When the vertex reaches λ_2 the α-cone is just tangent to the earth.

As λ is further increased, the α-trace moves across the globe in a generally equatorward direction. The maximum latitude $\lambda^\dagger$ attained by any part of the α-trace is given in Table 1. We see that for the first two orbits $\lambda^\dagger$ is greater than 60 degrees. However, the α-traces of points 5 and 6 never get within the source region, which is assumed to be confined to latitudes $>\lambda_C=60$ degrees. With this value of λ_C the α-trace for point 3 first enters the source region when $\lambda=11.0$ degrees and leaves the source region when $\lambda=13.4$ degrees. Thus, injection takes place only over a very small part of the total orbit (which mirrors at 32.2 degrees). The corresponding range for point 4 is 9.0 degrees to 9.7 degrees.

In comparing ϵ with the corresponding coefficients $\bar{\eta}_0$ calculated for the global component[9] it should be noted that the latter coefficient refers to an isotropic albedo whose strength is independent of latitude.

TABLE 3.—*Time-Integrated Solar Cosmic Ray Intensities Above* 10 *Mev**

Date	$J_o t_o$ particles/cm²-ster × 10^{-6}	Reference
February 23, 1956	1000	21
July 7, 1958	10	21
August 22, 1958	4	21
May 11, 1958	7	21
July 10, 1959	8	21
July 14, 1959	140	21
July 17, 1959	400	21
April 1-2, 1960	2	24
April 28-30, 1960	1	24
April 4-7, 1960	1	24
September 3, 1960	0.6	18
September 3-9, 1960	0.8	24
November 12-14, 1960	640	24
November 15, 1960	180	24
November 12-15, 1960	1000	19
November 20-26, 1960	20	24

*In each case extrapolations down to 10 Mev have been made assuming J_0 proportional to $T^{-1.3}$ except for the September and November, 1960, events where spectra directly observed in references 18 and 19 were used.

INTENSITY OF THE ANOMALOUS COMPONENT

Using the values $\sigma_0 = 80$ mb, $l = 15$ gm/cm², and $\Phi^* = 0.025$ (Model 3) we obtain

$$j(E,L,\alpha_o) = 6.4 \times 10^{-4}\, \epsilon(L,\alpha_o)\, S(E)\, J_o \frac{t_o}{t_n} \quad (35)$$

for the differential, directional intensity of newly trapped protons following an injection "impulse" of duration t_0. For the NERV points (ϵ on the order of 0.1) we therefore have an integral directional intensity $J(E)$ above 10 Mev on the order of $1.8 \times 10^{-8}\ J_0 t_0 p/\text{cm}^2\text{-ster-sec}$.

Now, the intensities observed at the NERV points (3 and 4) were approximately $10^3 p/\text{cm}^2$-ster-sec. In order to account for these intensities on the basis of impulsive injection from a single event we require $J_0 t_0$ to be approximately 5×10^{10}. The September 3, 1960, event (16 days prior to the date of the observations) was therefore far too small.

The events of July, 1959, contributed a time-integrated flux $J_0 t_0$ of approximately 5×10^8 while the large storm of February, 1956, reached approximately $10^9 p/\text{cm}^2$-ster. However, to evaluate the contribution from these less recent events we need to consider the lifetimes of the protons.

If protons are adiabatically trapped, then their lifetime, as determined by energy loss, is inversely proportional to the exospheric density averaged over the orbit. Assuming a hydrogen exosphere, 50 percent ionized, the $1/e$ lifetime, defined as the time over which the energy decreases to $1/e$ times the initial value, is given by

$$T_{1/e} = 10^{11} \frac{E^{1.5}}{\bar{\rho}} \text{ seconds}, \quad (36)$$

where E is the initial energy in Mev and $\bar{\rho}$ is the average density (atoms-and-ions/cm³). This expression is valid for $E < 300$ Mev. For orbits that mirror above approximately 1500 km $\bar{\rho}$ is less than $10^4/\text{cm}^3$. Therefore, if 10 Mev protons are adiabatically trapped their lifetime is 10 years or more.

The length of this lifetime suggests the possibility that an equilibrium exists, with injection from large events at irregular intervals of a few years being averaged over several events. Webber[21] has collected data on some twenty solar cosmic ray events since 1956. For most of these events the intensity only above about 100 Mev is known. We have extrapolated each spectrum to 10 Mev by assuming a spectrum J_0 proportional to $T^{-1.3}$ below 100 Mev (see Table 3).

We find that over a period of about 5 years

prior to the November 1960 events solar cosmic rays contributed a total of approximately 1.5×10^9 particles/cm^2-ster over the polar caps. This represents an average rate $\bar{J}$ of about 10 particles/cm^2-ster-sec, about 100 times the galactic rate.

An estimate of the equilibrium intensity from "continuous" polar injection is obtained by replacing J_0t_0 in equation 35 by $\bar{J}T_{1/e}$. Thus, this process can account for the observations provided $\bar{\rho}\approx10^3$/cm^3.

We therefore suggest that the "anomalous" component is a permanent feature. It might be called the polar component.

Since the injection is intermittent the polar component will behave as an equilibrium distribution only at energies that are high enough to make the lifetimes long compared to the interval between injection events. The lower end of this energy range will vary with altitude, i.e., with exospheric density, approximately as $r^{-2\nu/3}$ in a region where the density varies as $r^{-\nu}$.

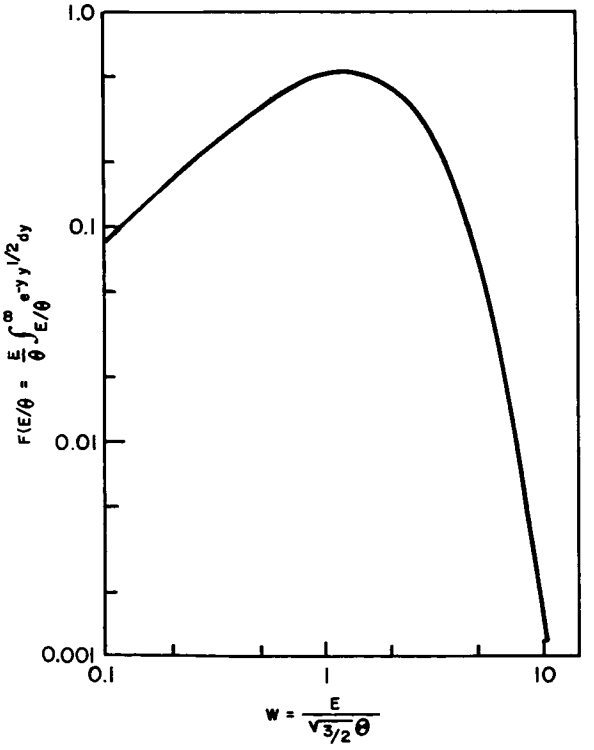


FIGURE 4.—The equilibrium energy spectrum resulting from injection with the Maxwellian spectrum and slowing down by pure energy loss is proportional to $F(w)$, where $E=\sqrt{\frac{3}{2}}\,\Theta w$.

In more detail, the equilibrium spectrum is found by integrating the continuity equation

$$\bar{q} = \frac{\partial}{\partial E}\left[n\frac{\partial E}{\partial t}\right], \tag{37}$$

assuming that v is proportional to $E^{1/2}$ and

$$\frac{\partial E}{\partial t} = -k\bar{\rho}E^{-1/2}, \tag{38}$$

where $k=7.65\times10^{-12}$ Mev$^{3/2}$-cm^3/sec for E in Mev. We find

$$j = \frac{\sigma_o \ell\, \Phi^* \Theta^{1/2}}{2\sqrt{\pi} k t_n}\frac{\epsilon\bar{J}}{\bar{\rho}}F = 1.7\times10^5\frac{\epsilon\bar{J}F}{\bar{\rho}}, \tag{39}$$

where $\epsilon=\epsilon(E,\alpha_0)$ and

$$F = F(E/\Theta) = \frac{E}{\Theta}\int_{E/\Theta}^{\infty} e^{-y} y^{1/2} dy. \tag{40}$$

The function F is related to functions tabulated by Pearson[25] and is shown in figure 4.

Before a comparison with the observations can be made we must take into account that, on the hypothesis of this paper, the spectrum at points 3 and 4 represents a superposition of a polar component upon the global component. We therefore take the spectrum seen at points 5 and 6, normalized to the data at 70 Mev, to be the global component and subtract this intensity from the anomalous spectrum.

The result for point 3 is shown in figure 5A along with the spectrum computed from equation 39 and normalized by setting $\bar{\rho}_3=1.06\times10^3$/cm^3. The corresponding result for point 4, using $\bar{\rho}_4=260$/cm^3, is shown in figure 5B. Noting that the neutron spectrum is only approximately given by the Maxwellian which we have assumed and that we have neglected the modulation of the albedo spectrum which result from energy degradation in the atmosphere we regard the agreement as at least fair. (The latter effect will tend to steepen the spectrum at the low energy end and to fill in the region below the peak in the source spectrum.)

An important point should be noted. The relative intensities at points 3 and 4 are about 1:3. That is, the higher intensity is found on the inner shell ($L=1.72$). Now the intensity ratio is predicted to be $\epsilon_3\bar{\rho}_4/\epsilon_4\bar{\rho}_3$, using subscripts to denote

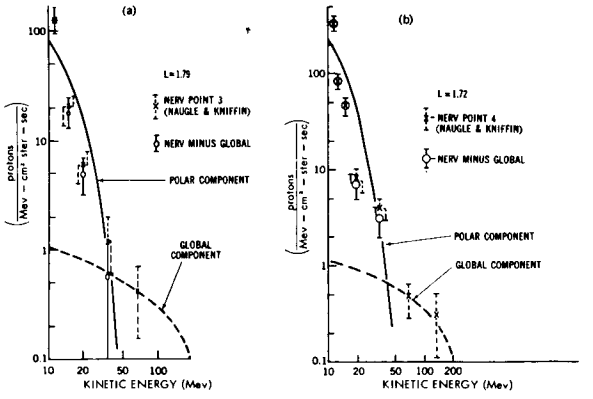


FIGURE 5.—Comparison between equilibrium spectrum and observations. The "global component" represents the spectrum observed at the innermost point, normalized to the data at 20 Mev. This component is subtracted from the observations (crosses) to give the anomalous component (circles). The heavy solid curve is the computed spectrum, normalized by the choice of density $\bar{\rho}$. In 5a, $L=1.79$, $\bar{\rho}=1.06\times10^3/\text{cm}^3$ and in 5b, $L=1.72$, $\bar{\rho}=2.6\times10^2/\text{cm}^3$. These intensities are computed with an assumed average solar cosmic ray flux $\bar{J}=10$ particles/cm²-ster-sec >10 Mev.

the number of the point. Since $\epsilon_3/\epsilon_4\approx1.5$ we require $\bar{\rho}_3/\bar{\rho}_4\approx4$. The average densities are proportional to the mirror point densities. Orbit 3 mirrors at 1600 km and orbit 4 mirrors at 1884 km. If the temperature is approximately 2000°K at the base of the exosphere then these mirror points will fall within the region in which oxygen predominates over hydrogen. The scale height of exospheric atomic oxygen at 1600 km is about 164 km for $T=2000$°K, leading to a ratio of mirror point densities $\exp[(1884-1600)/164]=5.6$. (If O^+ contributes strongly then this ratio will be reduced. A pure O^+ atmosphere leads to a ratio $\exp[284/(2\times164)]=2.4$.) We conclude therefore that the relative intensities at the NERV points 3 and 4 are consistent with our knowledge of the exosphere and do not indicate a breakdown of the trapping condition at these low energies (below 30 Mev).

SPATIAL DISTRIBUTION

The spatial distribution produced by this mechanism is characterized by two prominent features: (1) a *shadow zone* is formed close to the earth within which no "polar protons" are trapped, (2) beyond the shadow zone trapping is possible. However, there is a range of equatorial pitch angles centered about 90 degrees within which no injection takes place. This leads to the existence of a *penumbra*, a region in which the polar component is present at local pitch angles less than 90 degrees but does not mirror.

The shadow zone and penumbra formed by a polar cap source with boundary at $\lambda_C=60$ degrees are shown in figure 6. The shape of the penumbra is not a strong function of λ_C. Its boundary remains approximately parallel to the equatorial plane as λ_C is varied from about 30 to 90 degrees. The boundary of the shadow zone is always a magnetic shell of course.

PROTONS IN THE OUTER BELT

The polar mechanism injects about 100 times more low energy protons per Mev at about 10 Mev into the outer belt than does the global source. However, it is believed that the trapping ability of the quiescent field decreases at great altitudes. A variation of maximum trappable energy with altitude of the form $E_{max}=\text{constant}\times L^{-4}$ has been predicted by Singer.[26] Normalization of this relation is provided by satellite observations of the

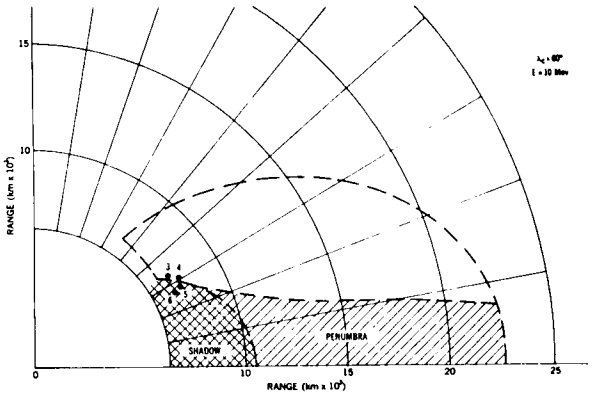


FIGURE 6.—Shadow zone and penumbra formed by polar cap injection. There is no trapping of fast protons at all in the shadow zone. Polar protons are trapped in the penumbra but they do not mirror in this region. The unshaded area may contain polar protons at all local pitch angles outside of the loss cone. The outer boundary of this region depends on energy as there is a maximum energy of proton which may remain trapped at a given L. The boundary drawn here corresponds to 10-Mev protons, assuming $E_{max}=1600L^{-4}$ Mev (reference 26). The points labeled 3, 4, 5, 6 are the locations of the NERV observations (reference 1).

radial extent of 75-Mev protons,[27] leading to $E_{max}=1600\ L^{-4}$ Mev. On this model, 1-Mev protons may remain trapped up to $L=6.3$, 10-Mev protons up to $L=3.6$, etc.

We may therefore expect to find the "polar" low-energy component extending to the heart of the outer zone or beyond. The intensity at a given energy should increase with altitude (as both ϵ and $1/\bar{p}$ increase) until E exceeds $E_{max}(L)$.

CONCLUSION

We have shown that the lower boundary of the proton belt produced by solar cosmic rays through the intermediary of albedo neutrons is consistent with the observed spatial distribution of low-energy protons in the inner zone. The energy spectrum of the high-latitude component is approximately accounted for (above 10 Mev) assuming secondary neutrons are generated in an evaporation spectrum and that the decay protons are in equilibrium. The absolute intensities roughly agree with experiment, assuming a reasonable source strength and atmospheric model.

In addition to the more obvious experimental tests such as extending the observations to lower energies and to higher L-values and such as repeating the same observations at a later time to look for time variations, we may point out that the polar source leads to a peculiar angular distribution (equatorial orbits are not populated) which can be investigated observationally (Lencheck and Singer, 1962).

ACKNOWLEDGMENTS

The author sincerely appreciates the close cooperation of J. Naugle and D. Kniffen. A part of this work has been submitted in partial fulfillment of the requirements for the PhD degree at the University of Maryland, thesis research under the direction of Prof. S. F. Singer and partially supported by a fellowship awarded by the National Rocket Club, Washington, D. C. The author is grateful to the National Aeronautics and Space Administration and the University of Maryland for their cooperation in this regard. Also, he acknowledges very helpful discussions with C. E. Fichtel and C. J. Waddington.

REFERENCES

1. Naugle, J. E., and Kniffen, D. A., "Flux and Energy Spectra of the Protons in the Inner Van Allen Belt," *Phys. Rev. Letters*, **7**(1): 3–6, July 1, 1961.
2. Freden, S. C., and White, R. S., "Particle Fluxes in the Inner Radiation Belt," *J. Geophys. Res.*, **65**(5): 1377–1383, May 1960.
3. Lenchek, A. M., and Singer, S. F., "Geomagnetically Trapped Protons from Cosmic-Ray Albedo Neutrons," *J. Geophys. Res.*, **67**(4): 1263–1287, April 1962.
4. Singer, S. F., "Trapped Albedo Theory of the Radiation Belt," *Phys. Rev. Letters*, **1**(5): 181–183, September 1, 1958.
5. Vernov, S. N., and Grigorov, N. L., et al., "Possible Mechanism of Production of 'Terrestrial Corpuscular Radiation' Under the Action of Cosmic Rays," *Doklady Akademii Nauk SSSR*, **124**(5): 1022–1025, February 11, 1959 (in Russian); Translation in *Soviet Phys.—Doklady*, **4**(1): 154–157, August 1959.
6. Armstrong, A. H., and Harrison, F. B., et al., "Charged Particles in the Inner Van Allen Radiation Belt," *J. Geophys. Res.*, **66**(2): 351–357, February 1961.
7. Lenchek, A. M., and Singer, S. F., "Injection of Trapped Protons from Solar Flare Particles," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961. II. *Joint Sessions*, Tokyo: Physical Society of Japan, 1962, pp. 123–127.
8. Wentworth, R. C., and Singer, S. F., "The Albedo Contribution in the Measurement of Cosmic-Ray Primaries," *Phys. Rev.*, **98**(5): 1546–1547, June 1, 1955 (Abstract).
9. Singer, S. F., "Latitude and Altitude Distribution of Geomagnetically Trapped Protons," *Phys. Rev. Letters*, **5**(7): 300–303, October 1, 1960.
10. Kellogg, P. J., "Electrons of the Van Allen Radiation," *J. Geophys. Res.*, **65**(9): 2705–2713, September 1960.
11. Hess, W. N., Canfield, E. H., and Lingenfelter, R. E., "Cosmic-Ray Neutron Demography," *J. Geophys. Res.*, **66**(3): 665–677, March 1961.
12. Tai, Y.-K., and Milburn, G. P., et al., "Neutron Yields from Thick Targets Bombarded by 18- and 32-Mev Protons," *Phys. Rev.*, **109**(6): 2086–2091, March 15, 1958.
13. Ajzenberg, F., and Franzen, W., "Neutrons from the Proton Bombardment of N^{11}," *Phys. Rev.*, **94**(2): 409–411, April 15, 1954.
14. Adelson, H. E., "Energy Spectra and Angular Dependences of Neutrons from the 31.5-Mev Proton Bombardment of Beryllium-9, Nitrogen-14 and Aluminum-27," Univ. Calif., Lawrence Radiation Lab., UCRL-8568, December 11, 1958.
15. Blatt, J. M., and Weisskopf, V. F., "Theoretical Nuclear Physics," New York: John Wiley and Sons, 1952, p. 350, equation (4.5).

16. LE COUTEUR, K. J., "The Evaporation Theory of Nuclear Disintegrations," *Proc. Phys. Soc. (London)*, **63**A(3): 259–282, March 1950.
17. ROSSI, B. R., "High-Energy Particles," New York: Prentice-Hall, 1952, p. 344.
18. DAVIS, L. R., FICHTEL, C. E., et al., "Rocket Observations of Solar Protons on September 3, 1960," *Phys. Rev. Letters*, **6**(9): 492–494, May 1, 1961.
19. OGILVIE, K. W., BRYANT, D. A., and DAVIS, L. R., "Rocket Observations of Solar Protons during the November 1960 Event," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961. II. *Joint Sessions*, Tokyo: Physical Society of Japan, 1962, pp. 317–319; also "Rocket Observations of Solar Protons During the November 1960 Events, I," *J. Geophys. Res.*, **67**(3): 929–937, March 1962.
20. FREIER, P. S., NEY, E. P., and WINCKLER, J. R., "Balloon Observation of Solar Cosmic Rays on March 26, 1958," *J. Geophys. Res.*, **64**(6): 685–688, June 1959.
21. WEBBER, W. R., "Time Variations of Low Energy Cosmic Rays during the Recent Solar Cycle," in: *Progress in Elementary Particle and Cosmic Ray Physics*, ed. by J. G. Wilson and S. A. Wouthuysen, Amsterdam: North-Holland Publ. Co., Vol. 6, 1962 (in press).
22. MCILWAIN, C. E., "Coordinates for Mapping the Distribution of Magnetically Trapped Particles," *J. Geophys. Res.*, **66**(11): 3681–3691, November 1961.
23. WENTWORTH, R. C., "Lifetimes of Geomagnetically Trapped Particles Determined by Coulomb Scattering," Ph.D. Thesis, University of Maryland, 1960.
24. LIN, W. C., "Observation of Galactic and Solar Cosmic Rays from October 13, 1959 to February 17, 1961 with Explorer VII (Satellite 1959 Iota)," State Univ. of Iowa SUI–61–16, August 1961 (Thesis submitted for M.S. degree).
25. PEARSON, K., "Tables of the Incomplete Gamma-Function," Cambridge: The University Press, 1957 (Reprint of 1922 edition).
26. SINGER, S. F., "Cause of the Minimum in the Earth's Radiation Belt," *Phys. Rev. Letters*, **3**(4): 188–190, August 15, 1959.
27. FAN, C. Y., MEYER, P., and SIMPSON, J. A., "Dynamics and Structure of the Outer Radiation Belt," *J. Geophys. Res.*, **66**(9): 2607–2640, September 1961.

IMPULSIVE INJECTION OF ELECTRONS INTO THE EARTH'S INNER TRAPPING REGION FROM SOLAR COSMIC-RAY EVENTS

MORDEHAI LIWSHITZ AND A. M. LENCHEK

Department of Physics and Astronomy, University of Maryland

The decay of albedo neutrons generated in the polar caps by solar high-energy-particle events provides an impulsive source of geomagnetically trapped electrons. The decay density due to slow albedo neutrons (which describe ballistic orbits above the atmosphere) can be calculated from simple kinematic considerations. At constant B the injection density exhibits a peak at intermediate L. At $L=2$ and $B=0.22$ gauss a typical large SHEP event deposits $\leq 3\times10^{-8}$ electron cm^{-3}. The flux from these electrons is $\sim10^{3}$ electrons cm^{-2} sec^{-1} in the intervals 0–780 kev. On account of the low efficiency for detecting these electrons, it appears that this process alone cannot account for the time variations seen in Explorer 7.

INTRODUCTION

Extended observations of the inner trapping region [Pizzella et al., 1962] show that the counting rates of detectors flown in the inner trapping region display noticeable increases following the occurrence of solar cosmic-ray events. Since no concurrent significant variations in the intensity of galactic cosmic radiation were observed during these periods, it appears plausible that the enhancement results directly or indirectly from the impact of solar particles upon the earth and its vicinity. Two mechanisms that may be responsible for this phenomenon are (a) the neutron albedo from solar protons striking the polar caps; (b) the injection of solar plasma into the magnetosphere and subsequent scattering into its inner regions.

Protons from decay of albedo neutrons produced by mechanism *a* may be ruled out as a source of the increased counting rates in the inner region since this region represents a 'shadow zone' for fast protons [Lenchek and Singer, 1962; Lenchek, 1962]. That is, at energies above a few tens of volts, there is no trapping for $L<1.6$. *Electrons*, however, differ from protons in this respect: their energy spectrum is fundamentally the same whether they are the decay product of low-energy neutrons or neutrons in the 10-Mev region. Consequently, electrons with energies of hundreds of kilo electron volts will be injected into orbits within the shadow zone, which is accessible to *slow* polar-cap neutrons, i.e., those with energies around 1 ev that are gravitationally 'trapped' and describe ballistic orbits.

Moreover, the fact that the amplitude of the enhancement of the counting rate increases with increasing L, where L is McIlwain's shell parameter [McIlwain, 1961], a priori agrees well with the polar-cap hypothesis, since higher L shells intersect the top of the terrestrial atmosphere closer to the source region.

The magnitude of this mechanism can easily be estimated. At the top of the atmosphere the decay density resulting from *galactic* cosmic-ray albedo neutrons is $\sim6\times10^{-12}$ decay cm^{-3} sec^{-1} at middle latitudes where the source strength is 5 neutrons per cm^2 column per second [Hess et al., 1961]. We may use the ratio of these two values to estimate the number of electrons injected by a solar cosmic-ray pulse.

The yield of neutrons per primary from the

solar event, Y_s, is approximately

$$Y_s = \frac{(\text{ionization range of primary}) \times (\text{neutrons/interaction})}{(\text{interaction mean free path})}$$

Because of the steep slope of the solar cosmic-ray spectrum the effective ionization range is ~ 1 g cm^{-2} (30 Mev). The interaction MFP is ~ 60 g cm^{-2} in this energy interval, and the average number of neutrons per interaction is $\sim \frac{1}{5}$. These values are consistent with the yield measured by Tai et al. [1958], who bombarded a thick nitrogen target with 31-Mev protons and measured a yield of 3.2×10^{-3} neutron per proton. The yield from oxygen is comparable. Therefore, $Y_s \sim 3 \times 10^{-3}$ neutron per primary.

Now, a very large solar cosmic-ray event deposits $\sim 10^9$ primaries on a square centimeter integrated over the duration of the event. This therefore produces 3×10^6 neutrons cm^{-2}. At a point at low altitude and high latitude *within the polar cap* the decay density will be very nearly what it would be if the solar cosmic rays were incident upon the entire globe. Using the figures mentioned above, the time-integrated decay density is

$$3 \times 10^6 \times 6 \times 10^{-12}/5 = 3.6 \times 10^{-6} \quad \text{decay cm}^{-3}$$

The flux due to the β-decay electrons is therefore

$$\sim 3.6 \times 10^{-6} \times 3 \times 10^{10} \simeq 10^5 \quad \text{cm}^{-2}\,\text{sec}^{-1}$$

if all are trapped. When we take into account the flight time, and the geometry of the ballistic orbits, this value is reduced for low altitudes on shells with $L \sim 2$ by about two orders of magnitude, one owing to decay in flight, and one in view of the distance between the point of observations and the finite source region. At such a location the decay electron flux is therefore $\sim 10^3$ cm^{-2} sec^{-1}.

SPATIAL DISTRIBUTION OF NEUTRON DECAY DENSITY

We wish to evaluate the spatial distribution of neutrons emerging from a localized source. The neutrons describe ballistic orbits above the atmosphere without collisions and form a neutron exosphere. As an initial step in this direction we may start with the distribution of neutrons from a point source at the geomagnetic pole. As will be seen, this is a satisfactory approximation for our purpose if we confine our attention to moderate L values, the corresponding shells being fed by neutron decay not too close to the source region. (The source region will later be assumed to extend from the poles to 60° latitude.) The main constituents of the neutrons in the low-energy range of interest to us in the inner trapping region are diffusion neutrons, for which the assumption of isotropic emission introduces no serious error. Let $f(v)$ sec^{-1} (cm/sec)$^{-1}$ be the total neutron albedo flux from the point source at the pole in the velocity interval $(v, v+dv)$; let R_E be the earth radius; r, the geocentric distance of a point in space; $R = r/R_E$; $\bar{\lambda}$, its colatitude. Let w and ϕ be the velocity and zenith angle of a neutron at $(r, \bar{\lambda})$ that started at the pole with v and θ. Let t' be the time of neutron flight from point source to field point, and let $t_n = 10^3$ sec be the neutron lifetime. Then the contribution to the differential neutron density $N(v, r, \bar{\lambda})$ from such particles may be written

$$N(v, r, \bar{\lambda})\, dv = f(v)\, \sin\theta\, d\theta (2\pi r\, \sin\bar{\lambda} w\, \sin\phi\, dr)^{-1} \cdot \exp -(t'/t_n)\, dv \quad (1)$$

This expression equates the normal component of the flux through a ring of radius $r \sin\phi$ and width dr at colatitude $\bar{\lambda}$ to the outflux of corresponding particles from the source. From conservation of angular momentum

$$rw \sin\phi = R_E v \sin\theta \quad (2)$$

whence (1) becomes

$$N(v, R)\, dv = f(v)\, (2\pi R_E^2 \sin\bar{\lambda} v)^{-1}\, (dv/dR) \cdot \exp -(t'/t_n)\, d\theta \quad (3)$$

Now, motion under a central force such as gravity defines a unique relation between v and θ for given R. The right side of equation 3 may be expressed in terms of θ alone, and may then be

integrated to yield the integral neutron density $\mathfrak{N}(R, \bar{\lambda})$, i.e., neutrons cm^{-3}.

This, however, entails knowledge of the neutron albedo spectrum $f(v)$. The present study has utilized the neutron flux investigated by Hess et al. [1961] and extrapolated by them to the top of the atmosphere on the basis of diffusion theory. We are primarily concerned with the low-velocity region of the spectrum. This region is not affected by the primary spectrum. We are therefore justified in assuming that the slow neutron albedo flux from a polar event differs from the ambient galactic slow neutron albedo flux in magnitude only, but not in shape. The flux is therefore of the form

$$i_n(v)\ dv = (A/v)\ \exp(-b/v)\ dv = Ah(v)\ dv \quad (4)$$

A good fit to the curve of Hess et al. [1961] for 0 g cm^{-2} gives for the albedo flux due to galactic cosmic rays

$$i_n(v) = 7.5\times10^{-8}(1/v)\ \exp(-6/v) \times \text{neutrons cm}^{-2}\ \text{sec}^{-1}\ (\text{cm/sec})^{-1} \quad (5)$$

where v is in kilometers per second.

A varies from event to event and also changes during each event. We may arrive at a value for it in the following way. The total neutron concentration in the atmosphere is proportional to the source strength. This holds for both the solar and the ambient galactic components. In both cases, therefore,

$$\mathfrak{N} = A\int_0^\infty \left[\frac{h(v)}{v}\right] dv \quad (6)$$

Since the bulk of the neutron density is contributed by the peak in the velocity spectrum at about thermal velocity, no appreciable error is introduced by extending the integration to infinity in the case of solar neutron albedo as well as galactic neutron albedo. It follows that the coefficients A are directly proportional to the source strength S, the number of neutrons generated cm^{-2} sec^{-1}. Thus

$$A_s = A_g(S_s/S_g) \quad (7)$$

where the subscripts g and s refer to the galactic and solar component, respectively. S_s can be approximated as follows: Let $Y(T)$ neutrons/proton be the neutron yield from a proton incident at kinetic energy T, and let $i_p(T)$ be the differential omnidirectional primary proton flux cm^{-2} sec^{-1} Mev^{-1}. Then

$$S_s = \int_{T_1}^{T_2} Y(T) i_p(T)\ dT \quad (8)$$

T_1 and T_2 are the respective limiting energies of the solar proton spectrum. A rough measure of $Y(T)$ is given by

$$Y(T) = R(T)/l \quad (9)$$

as was mentioned in the introduction, where $R(T)$ is the proton ionization range in air [$\sim 3\times10^{-3}T^{1.7}$ (Mev) g cm^{-2}] and l is the mean neutron production range, assumed constant over energy $\sim 3\times10^2$ g cm^{-2}.

With $f(v)$ calculated in the indicated manner, we can then integrate equation 3. As can be shown from the expressions derived in appendix A, equation 3 assumes a relatively simple form: in the equatorial plane, and along the top of the atmosphere, at various colatitudes. These expressions are given respectively by

$$\mathfrak{N}(R, \bar{\lambda}=\pi/2) = 2\times0.85\times10^{-2}A_sR^{-2}\int_{\theta_{\min}}^{\pi/2}\left\{\sin^2\theta\left(\frac{\sin^2\theta}{R}+\sin\theta\cos\theta\right)^{-1/2}\exp(-t'(\theta)/t_n) \times \exp[-0.762\ (\sin^2\theta/R+\sin\theta\cos\theta)^{1/2}]\right\} d\theta \quad (10a)$$

$$\mathfrak{N}(R=1, \bar{\lambda}) = 0.85\times10^{-2}A_s(1-\cos\bar{\lambda})^{-1/2}\ (\sin\bar{\lambda})^{-1}\int_{\theta_{\min}}^{\pi/2}\{\sin^2\theta\ [\sin^2\theta(1-\cos\bar{\lambda})+\sin\theta\cos\theta\sin\bar{\lambda}]^{-1/2} \times \exp[-0.762\ (\sin^2\theta\{1-\cos\bar{\lambda}\}+\sin\theta\cos\theta\sin\bar{\lambda})^{-1/2}(1-\cos\bar{\lambda})^{-1/2}]\exp[-t'(\bar{\lambda}, \theta)/t_n]\}\ d\theta \quad (10b)$$

plus a similar expression with $(\pi-\bar{\lambda})$ substituted for $\bar{\lambda}$, giving the contribution from the opposite pole.

Actual, though cumbersome, numerical integration shows that for moderate latitudes, say up to 45°, the density at the top of the atmosphere may be conveniently approximated by

$$\mathfrak{N}(R=1,\ \bar{\lambda}) = B[\bar{\lambda}^{-2.8} + (\pi-\bar{\lambda})^{-2.8}] \qquad (11)$$

Thus, we can use this as a Green's function, and integrate over the source distribution on the polar cap, which we assume to be uniform. It turns out, however, that even for $\bar{\lambda}$ as small as 45° the density calculated on the assumption of a point source differs by less than 50 per cent from that based on a source distributed over the polar cap. In view of the uncertainties in the many other parameters this suggests the use of the simpler model.

We turn now to the problem of the injection into trapped orbits. The neutron decay density (or equivalently electron injection density) is

$$Q = \mathfrak{N}/t_n \quad \text{cm}^{-3}\,\text{sec}^{-1} \qquad (12)$$

We may assume that the lifetimes of the newly injected electrons are appreciably longer than the duration of injection, an assumption supported by the data of Pizzella et al. [1962]. The data indicate lifetimes of the order of weeks and months. This serves as a justification for the assumption that (12) integrated over the duration of the event, yielding the total number injected, actually represents the total concentration of electrons present at the termination of the event.

By proceeding in the manner outlined above, the electron *flux*, anywhere at the time of the event's termination, can be evaluated, in principle.

The following notation is used:

α = the local pitch angle of electrons.

T' = their kinetic energy.

$N_\beta(T')$ = the differential electron spectrum normalized to unity.

j = the directional electron intensity (with $j_\perp$ the intensity normal to the magnetic lines of force).

ds = the infinitesimal spiral path length.

ν = the frequency of electron oscillation between mirror points.

$\oint ds$ = the integration over a complete bounce cycle.

$\mathfrak{N}_{\beta\,\text{tot}}(\alpha=\pi/2,\ R,\ T',\ L)$ = the total number kev^{-1} cm^{-2} ster^{-1} of electrons lying within the orbit $\oint ds$ mirroring at R, with energy T'.

t_0 = duration of injection event.

R_m = mirror distance in units of earth radii.

Then

$$\begin{aligned} j(L,\ R,\ \alpha=\pi/2,\ T') &= j_\perp(L,\ R,\ T') \\ &= \mathfrak{N}_{\beta\,\text{tot}}(L,\ \alpha(R)=\pi/2,\ T')\,\nu \\ &= \oint ds \frac{Q(s)}{4\pi} t_0 N_\beta \beta c \Big/ \oint ds \\ &= (t_0 N_\beta \beta c/4\pi) \oint Q(s)\, ds \Big/ \oint ds \\ &= (t_0 N_\beta \beta c/4\pi)\,\bar{Q} \end{aligned} \qquad (13)$$

where $\bar{Q}$ is the electron injection density averaged over the spiral path. To evaluate $\bar{Q}$ we need to know the altitude dependence of the decay density. Both the earlier results of Hess et al. [1961] and the results of this analysis are consistent with a decay density varying as $R^{-2.7}$ for moderate latitudes. It can be shown (see appendix B) that, if

$$Q(R) = Q(R_m)\,(R/R_m)^{-\gamma}$$

where R_m is the geocentric mirror point distance, then to a good approximation

$$\bar{Q} = Q(R_m)\frac{\Gamma(\gamma-\frac{1}{2})}{\Gamma(\gamma)}\left[\frac{\pi X_m^2(4-3X_m)^2}{6S^2(8-5X_m)(1-X_m)}\right]^{1/2} \qquad (14)$$

Here, $X_m = R_m/LR_E$ and

$$S = \frac{1}{4}\oint ds/LR_E \quad \text{(dimensionless)}$$

To obtain the omnidirectional flux we must integrate over local pitch angles, making use of

Liouville's theorem and the adiabaticity of the motion.

In view of equation 11 we can then represent the directional electron intensity at any point in a shell with parameter L as

$$j(L, \lambda, \alpha) = C[\lambda_m{}^{2.8}(L \sin^2\lambda_m)^{2.7}]^{-1}\frac{\Gamma(\gamma-\frac{1}{2})}{\Gamma(\gamma)} \cdot\left[\frac{X_m{}^2(4-3X_m)^2}{6S^2(8-5X_m)(1-X_m)}\right]^{1/2} \quad (15)$$

When viewed in the equatorial plane the initial distribution has a sharp peak at the pitch angle α_0 given approximately by

$$\sin^2\alpha_0 = L^{-3}(4-3/L)^{-1/2}$$

The peak has an amplitude nearly proportional to L for small L and proportional to $1/L$ for $L \gtrsim 8$. This behavior is easily understood. In averaging Q over the trapped orbit we are, in effect, dividing the 'scale height' of the decay density by the length of the line of force. For shells that intersect the earth outside the polar cap ($L \lesssim 4$) the decay density near the earth increases with L somewhat faster than linearly versus L. However, for shells that intersect well within the polar cap ($L \gtrsim 8$) the decay density at a constant low altitude is independent of L. In this case, referring to equation 14, $Q(R_m)$ becomes a constant and setting $X_m \simeq 1/L \ll 1$ yields

$$\bar{Q} \simeq \text{constant}/L$$

Therefore, the peak in the *directional* intensity immediately after the injection period displays a maximum at an intermediate value of L. The exact value of L on which the maximum is found cannot be determined here since the source distribution is not sufficiently well known.

The omnidirectional intensity is proportional to the width of the angular distribution as well as to the amplitude. For points at constant altitude h this width is proportional to approximately

$$1-\left(\frac{3h/2R_E}{4L-3}\right)^{1/2}$$

Thus, for h small compared with the earth radius R_E, the width is relatively independent of L, giving an omnidirectional intensity proportional to the peak directional intensity, i.e., varying about as L^{+1} for $L \lesssim 4$ and as L^{-1} for L somewhat greater than 4.

DISCUSSION OF NUMERICAL RESULTS

The expressions above do not lend themselves to integration in closed form and have to be integrated numerically. The constant A_s in equation 7 has been computed for a few selected events included in Bailey's [1962] review of recent solar events. The events chosen are: (a) 'typical event'

TABLE 1.—*Peak Omnidirectional Electron Fluxes for Invariable B (0.22 gauss) and Different L*

Type of event	A_s, cm^{-2} sec^{-1}	$I(L, B=0.22)$, cm^{-2} sec^{-1}		
		$L=1.3265$	$L=1.656$	$L=2.000$
'Typical'*	1.15×10^{-6}	165	420	570
Sept. 3, 1960*	5.65×10^{-8}	8.25	21.0	28.5
Nov. 12, 1960*	3.8×10^{-6}	550	1150	1850
Nov. 1960†	—	100–200	1000	1500
Sept. 3, 1960†	—	—	—	—
Galactic‡	7.5×10^{-8}			

*Computation based on data given in Bailey [1962].
†Observed by Pizzella et al. [1962].
‡Computation based on the work of Hess et al. [1962].
Note that only the additional flux in Pizzella's data has been taken into account in the compilation of this table.

(curve 2 in Figure 1 of this reference); (b) September 3, 1960 (curves 8 and 9); (c) November 12, 1960 (curve 10). Resulting values for A_s are included in Table 1.

Table 1 also gives the omnidirectional electron fluxes for locations characterized by the following values of the parameters B and L: $B=0.22$ gauss; $L=2.000$, 1.656, 1.3265. These particular values were chosen for ease of comparison with the experimental data of Pizzella et al. [1962], which were reduced by those authors to similar values of the parameters. For the evaluation of the corresponding R and λ, the real terrestrial field has been used [Jensen et al., 1960]. In each case the value of R_m chosen is the lowest point on the 'mirror trace' over longitude, the main contribution to the flux of electrons trapped in a particular magnetic shell coming from injection in the vicinity of the lowest mirror point. It has been assumed, furthermore, that electrons mirroring below altitude of 300 km are 'catastrophically' removed by atmospheric scattering. This effectively determines the 'loss cone' in the angular distribution, and thus takes partially into account particle losses sustained during the finite injection period. The exact value of the cutoff altitude does not materially affect the nature of the results, the main contribution to the omnidirectional flux at R originating within a region of a few hundred kilometers below R. As the lowest possible value of the cutoff altitude we may consider the 'top of the atmosphere' at about 100 km.

Comparison of the experimental and theoretical values of the electron flux shows good agreement for the events and locations chosen, *under the assumption that the bulk of the 'polar-cap electrons' are detected.* However, the detector of Pizzella et al. [1962] is similar to a detector flown in Explorer 12 which responds to β-decay electrons with an average efficiency about 10^{-5} [O'Brien et al., 1962]. Therefore, if these electrons are to produce the observed enhancement in the counting rate, they must undergo an acceleration raising their mean energy by a factor of about 3. It appears that such an acceleration may indeed occur. Pizzella's data reveal that the counting rates may lag the solar event by a number of days, of the order of a week. This, then, suggests that an acceleration is operative for some time after the injection ceases.

If the neutron decay process is unable to explain time variations accompanying SHEP events, we must look elsewhere. Neutron decay represents an *indirect* means of injecting *new* electrons (and protons). Direct injection, as from solar plasma, might fortuitously show a spatial dependence similar to that of the polar neutron source. That is, for low L both sources increase with increasing L. Further, injection taking place at the outer boundary of the trapping region can also lead to a transient peak in density at intermediate L [Parker, 1960]. Both direct and indirect injection introduces new particles in a number unrelated to the number of 'old' particles present before the event. Therefore, the *percentage* enhancement depends upon the prestorm level of intensity. On the other hand, acceleration of 'old' electrons by magnetic fluctuations or waves propagating through the magnetosphere in association with the SHEP event would produce an enhancement that might be proportional to the number of 'old' electrons at a given L. The existence of the artificial radiation belt may eventually make it possible to distinguish between injection of new electrons and acceleration of old. Since the number of electrons at low L is now considerably increased over what it was at the time of Explorer 7 observations, a large SHEP event now accompanied by proportionately smaller percentage of enhancement would argue in favor of injection (direct or indirect) and a similar percentage of enhancement would argue in favor of acceleration. Unfortunately, no large SHEP events have been reported since the creation of the artificial belt.

ACKNOWLEDGMENTS

This investigation was suggested by S. F. Singer. We also wish to acknowledge helpful discussions with H. J. Laster.

The work is supported by AFOSR grant AF-AFOSR-62-284.

Appendix A

EXPLICIT EXPRESSIONS FOR THE SPATIAL DISTRIBUTION OF NEUTRONS

The expressions appearing in equations 3 and 10 will now be derived explicitly. The following notation will be used:

$V_\infty = (2GM/R_E)^{1/2}$ escape velocity of a neutron at the earth's surface.

E = total energy of a neutron.

$u = v^2/v_\infty{}^2$.

ϵ = eccentricity of the neutron's orbit.

Other symbols have been defined in the text.

The orbital equation of a particle moving in an inverse square central field, such as the earth's gravitational field, may be written

$$1/r = (1-\epsilon\cos\psi)/a(1-\epsilon^2) \qquad \text{(A1)}$$

In bound elliptic orbits a is the semimajor axis. Consider a neutron emitted at the pole with velocity V and zenith angle θ. Then

$$a = -GmM/2E = GM/(v_\infty{}^2 - v^2) = 1/[2(1-u)] \qquad \text{(A2)}$$

Similarly, using some algebra, we obtain

$$1-\epsilon^2 = 4(u-u^2)\ \sin^2\theta \qquad \text{(A3)}$$

Then, combining A1, A2, and A3, we obtain

$$\epsilon\cos\psi = 1-2u\ \sin^2\theta \qquad \text{(A4)}$$

For a particle at r and colatitude $\bar{\lambda}$

$$1/r = (1-\epsilon\cos\psi\cos\bar{\lambda} - \sin\psi\sin\bar{\lambda})/a(1-\epsilon^2) \qquad \text{(A5)}$$

From the above equation, for fitted θ, the relation between u and r follows:

$$u = (1-\cos\bar{\lambda})/2[(1/r-\cos\bar{\lambda})\ \sin^2\theta + \sin\theta\cos\theta\sin\bar{\lambda}] \qquad \text{(A6)}$$

Equation A6 permits the evaluation of $(1/v)\cdot(dv/dR)$ in equations 3 and 10.

Appendix B

We wish to evaluate

$$\bar{f}=\oint f\,ds \Big/ \oint ds$$

for the case where f varies as $f(r)=\text{constant}\ r^{-\gamma}$. If we put $x=r/r_e$ and $\oint ds=4S$, with S in dimensionless units of $r_e=LR_E$,

$$\bar{f}=\frac{1}{S}\int_{x_m}^{1} f(x)\frac{ds}{dx}dx$$

where

$$\frac{ds}{dx}=(4-3x)^{1/2}/2(1-x)^{1/2}$$

$$\cdot\left[1-\left(\frac{x_m}{x}\right)^3\left(\frac{4-3x}{4-3x_m}\right)^{1/2}\right]^{1/2}$$

This expression for ds/dx follows directly from the equation of a dipole line of force, $r=r_e\cos^2\lambda$.

We expand ds/dx in a Taylor series about x_m, and put

$$f(x)=f(x_m)\,(x/x_m)^{-\gamma}$$

After some algebra and a change of variable we obtain

$$\frac{\bar{f}}{f(x_m)}=\left[\frac{x_m^2(4-3x_m)^2}{6S^2(8-5x_m)\,(1-x_m)}\right]^{1/2}$$

$$\cdot\int_0^{1/x_m}(1+w)^{-\gamma}w^{-1/2}\,dw$$

neglecting terms of order x_m and higher. For $\gamma\gtrsim 2$ the major contribution to this integral comes from the region near $w=0$. Therefore we approximate the integral by

$$\int_0^{\infty}(1+w)^{-\gamma}w^{-1/2}\,dw=\sqrt{\pi}\frac{\Gamma(\gamma-\frac{1}{2})}{\Gamma(\gamma)}$$

and obtain the result

$$\bar{f}\simeq f(x_m)\frac{\Gamma(\gamma-\frac{1}{2})}{\Gamma(\gamma)}\left[\frac{\pi x_m^2(4-3x_m)^2}{6S^2(8-5x_m)(1-x_m)}\right]^{1/2}$$

This result can be used to approximate $\bar{Q}$, since for orbits with small equatorial pitch angle the bulk of the injection takes place near the mirror point. We also assume that, since the line of force makes a relatively small zenith angle near the mirror point, the variation of Q with latitude can be neglected.

REFERENCES

Bailey, D. K., Time variations of the energy spectrum of solar cosmic rays in relation to the radiation hazard in space, *J. Geophys. Res.*, **67**, 391–396, 1962.

Hess, W. N., Canfield, E. H., and Lingenfelter, R. E., Cosmic-ray neutron demography, *J. Geophys. Res.*, **66**, 665–677, 1961.

Jensen, D. C., Murray, R. W., and Welch, J. A., Jr., Tables of adiabatic invariants for the geomagnetic field 1955.0, *AFSWC-TN-60-8*, 1960.

Lenchek, A. M., On the anomalous component of low-energy geomagnetically trapped protons, *J. Geophys. Res.*, **67**, 2145–2157, 1962.

———, and Singer, S. F., Geomagnetically trapped protons from cosmic-ray albedo neutrons, *J. Geophys. Res.*, **67**, 1263–1287, 1962.

———, ———, and Wentworth, R. C., Geomagnetically trapped electrons from cosmic-ray albedo neutrons, *J. Geophys. Res.*, **66**, 4027–4046, 1961.

McIlwain, C. E., Coordinates for mapping the distribution of magnetically trapped particles, *J. Geophys. Res.*, **66**, 3681–3691, 1961.

O'Brien, B. J., Laughlin, C. D., Van Allen, J. A., and Frank, L. A., Measurements of the intensity and spectrum of electrons at 1000-kilometer altitude and high latitudes, *J. Geophys. Res.*, **67**, 1209–1225, 1962.

Parker, E. N., Geomagnetic fluctuations and the form of the outer zone of the Van Allen radiation belt, *J. Geophys. Res.*, **65**, 3117–3130, 1960.

Pizzella, G., McIlwain, C. E., and Van Allen, J. A., Time variations of intensity in the earth's inner radiation zone, October 1959 through December 1960, *J. Geophys. Res.*, **67**, 1235–1253, 1962.

Tai, Y-K., Millburn, G. P., Kaplan, S. N., and Moyer, B. J., Neutron yields from thick targets bombarded by 18- and 32-Mev protons, *Phys. Rev.*, **109**, 2086, 1958.

COSMIC RAY EXPERIMENTS FOR EXPLORER XII (1961υ) AND THE ORBITING GEOPHYSICAL OBSERVATORY

GEORGE H. LUDWIG AND FRANK B. McDONALD

Goddard Space Flight Center

The cosmic ray experiment on Explorer XII consisted of a Geiger counter telescope, a thin CsI scintillation counter, and a large-area scintillation counter telescope. The thin scintillation counter was connected to an 8-level integral analyzer. The large-area scintillation counter telescope, which measured the energy loss of the detected particle, was fed to a 32 channel differential pulse height analyzer with a storage capacity of 65,535 counts per channel. Both the Geiger counter telescope and single counter rates were telemetered. All information was multiplexed onto a single channel. Details of the instrumentation and the methods of encoding are discussed therein.

For the first Orbiting Geophysical Observatory mission, a new scintillation counter telescope has been developed which measures both energy loss and total energy. This furnishes excellent charge and energy resolution over an energy range from about 11 to 90 Mev per nucleon. Design details of this telescope and its associated electronic instrumentation are also presented.

INTRODUCTION

Measurements of the primary cosmic ray charge and energy spectra have been conducted for some time by means of balloon-borne instruments and nuclear emulsions. These measurements have been confined to the energy range above 65 Mev/nucleon since the lower energy particles cannot penetrate the atmosphere above the balloon, and thus cannot reach the detectors. Rocket probes could carry the instruments farther from the earth to extend these investigations to lower energies; however, the cosmic ray flux is so low that adequate data could not be accumulated during the brief trajectory to yield statistically meaningful results. The high apogee satellite offers the first good chance to investigate the lower energy region of the cosmic ray spectrum. Such a satellite must place the instruments well outside the earth's high intensity radiation belts for an appreciable period of time.

THE EXPLORER XII EXPERIMENT

Explorer XII (1961υ) was launched on August 16, 1961, into an orbit having an initial geocentric apogee of 83,600 km and perigee of 6700 km. This satellite contained a detector to investigate the energy spectrum of the total primary cosmic ray flux in the range from 3 to greater than 600 Mev/nucleon. It consisted of three detector arrays: a scintillation detector employing a thin crystal to investigate the spectrum in the 3 to 80 Mev/nucleon range; a double scintillator telescope array to perform a detailed differential analysis of the 100 to 600 Mev/nucleon portion; and a Geiger-Mueller counter telescope to measure protons having energies greater than about 28 Mev.

The Single Scintillation Detector

The configuration of the thin, single crystal detector is shown in figure 1. A CsI(Tl) crystal 0.5 gm/cm^2 thick was used. A collimator was located around the crystal to define the geometry for all protons having energies less than about 400 Mev. The crystal was viewed by an RCA type C–7151 photomultiplier (PM) tube, which is a ruggedized version of the RCA 6199 tube. The detector was calibrated continuously by a small Pu^{239} alpha particle source mounted on the crystal; these

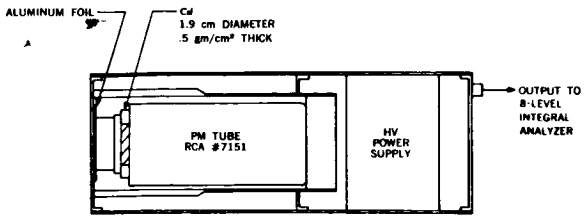


FIGURE 1.—The thin scintillator counter assembly.

particles lose about 3.5 Mev in the crystal. The high voltage power supply was located at the base of the PM tube. The complete assembly was encapsulated to form a rugged assembly capable of withstanding the launch environment and of operating throughout the pressure transition from one atmosphere to vacuum.

The characteristics of the detector are shown in figure 2. Protons which enter the crystal through the thin foil (6.5 mg/cm² aluminum) and lose all their energy in the crystal fall along the curve A. The point of intersection of curves A through the crystal but lose all their energy in the process. More energetic protons fall along curve B. As the particles become more and more energetic, less and less energy is lost due to the decreasing rate of energy loss. Minimum ionizing protons fall on the high energy, zero slope portion of the curve.

An eight-level integral pulse height analyzer was used with the detector; the pulse height thresholds are indicated in figure 2. For each setting, all events resulting in an energy loss in the crystal greater than the threshold value were

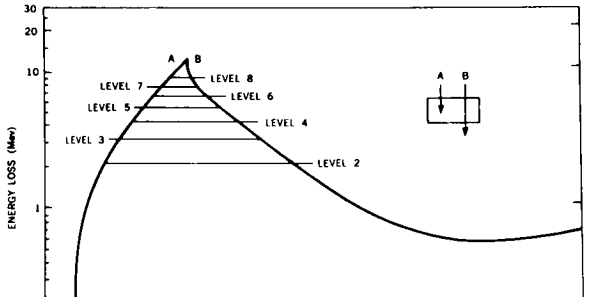


FIGURE 2.—Proton energy loss in the thin scintillator as a function of energy.

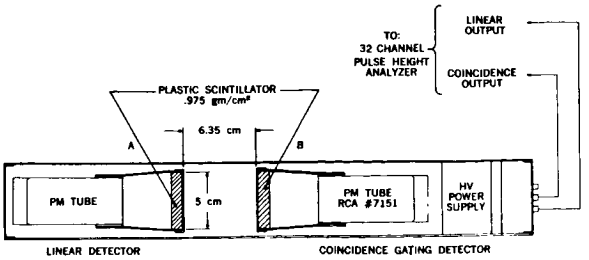


FIGURE 3.—Outline drawing of the scintillator telescope assembly.

counted. Thus, all events counted while the analyzer was on level eight corresponded to protons of 9.4 to 13 Mev, all events counted on level seven to protons of 8 to 14 Mev, and so on. From a combination of this information with that from the other detectors and our previous knowledge of the spectrum, it is possible to determine the spectral distribution of particles down to approximately 3 Mev per nucleon.

The Double Scintillator Telescope

Since the amount of absorbing material needed to define the geometry for higher energy particles is excessive for use in satellites, a double scintillator telescope array (figure 3) was employed for these particles. Two 0.975 gm/cm² disc shaped

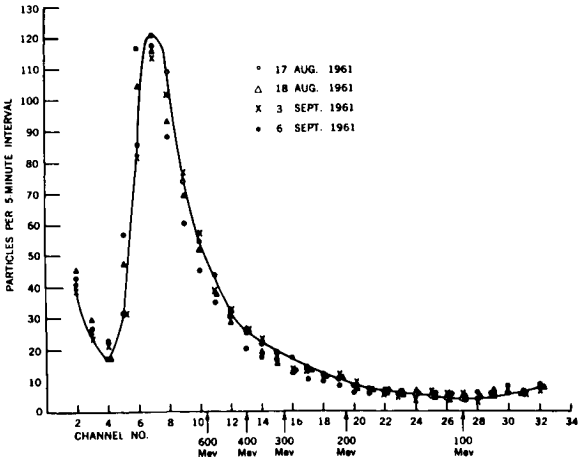


FIGURE 4.—Scintillator telescope pulse height distributions obtained on 4 days during the flight of Explorer XII. The energies of protons which pass completely through the crystal and the analyzer channel number are superimposed on the abscissa.

plastic scintillators were used in a telescope array having a geometric factor of 13.6 cm²-ster. The pulse from scintillator B was analyzed when a particle passed through scintillator A and into or through scintillator B. A curve similar to the one (figure 2) for the single scintillator existed for this B scintillator—the protons which entered but did not penetrate the scintillator fell along an ascending curve A, and those which penetrated fell along a curve B.

The analysis of these events was performed by a 32 channel differential pulse height analyzer with a magnetic core memory. Figure 4 shows a spectrum obtained from this instrument during its flight. A large portion of the events (during solar quiet times) are minimum ionizing protons corresponding to an energy loss of about 1 Bev—thus, we find a peak in the spectrum at that energy which assists in calibrating the system. This peak is spread by the Landua spread of the minimum ionizing particles. The events registered in channels 2 through 4 are due to low energy particles. The failure of the curve to terminate at an energy loss corresponding to the intersection of curves A and B in figure 2 is due to the alpha particle contribution. For this reason the useful lower energy limit of the detector is set at about 100 Mev. The upper limit is about 600 Mev, since protons of greater energy exhibit large statistical fluctuations in energy loss.

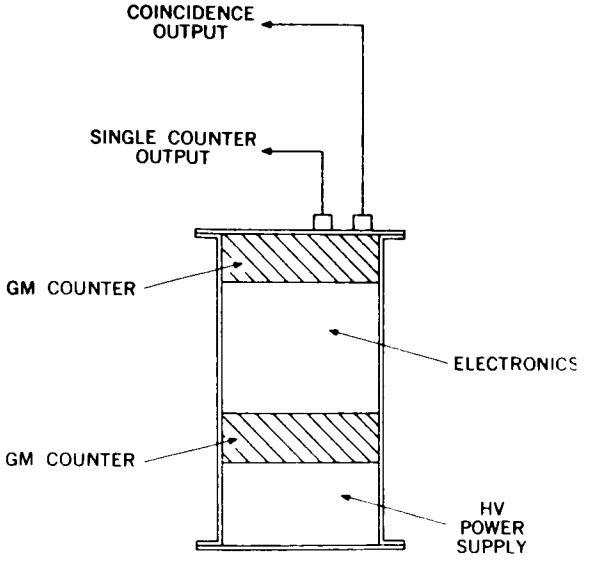


FIGURE 5.—The GM counter telescope configuration.

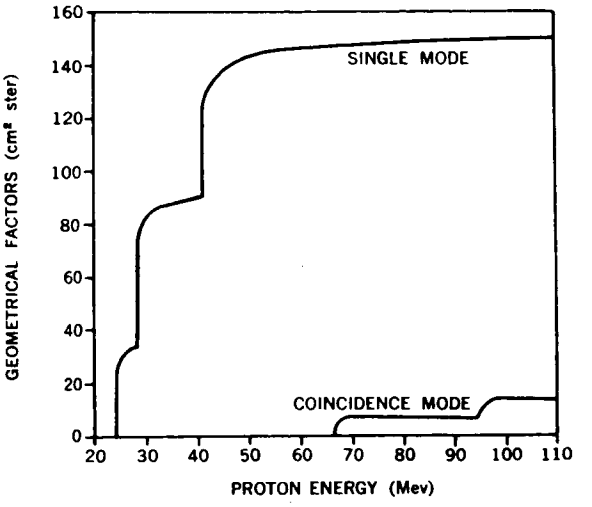


FIGURE 6.—Geometric factor as a function of incident proton energy for both modes of operation of the GM counter telescope. The geometry is defined by a combination of the counter configuration and the absorbing material in the proton path.

The GM Counter Telescope

A simple GM counter telescope was flown to determine the directional and omnidirectional cosmic ray intensities. The telescope threshold energies were 70 and 8 Mev for protons and electrons, respectively. The GM counters were disc shaped with effective diameters of 4.45 cm and depths of 1.0 cm. Figure 5 shows the configuration of the array, and the geometric factors are plotted as a function of proton energy in figure 6. The efficiency of the telescope was 88 percent for minimum ionizing protons. It was operated in two modes: In the coincidence (telescope) mode, only particles entering both counters were counted. In the singles mode, all events from one GM counter were counted to obtain an approximation of the omnidirectional flux.

The Signal Conditioning Instrumentation

A signal conditioning system[1] processed the signals from the three detector arrays (figure 7). The pulses from the single GM counter, the coincidence pulses from the GM counter telescope, and the thin scintillator pulses from the eight-level threshold discriminator were counted sequentially by a 15-stage binary scaler. At the beginning of the subcommutation cycle of the

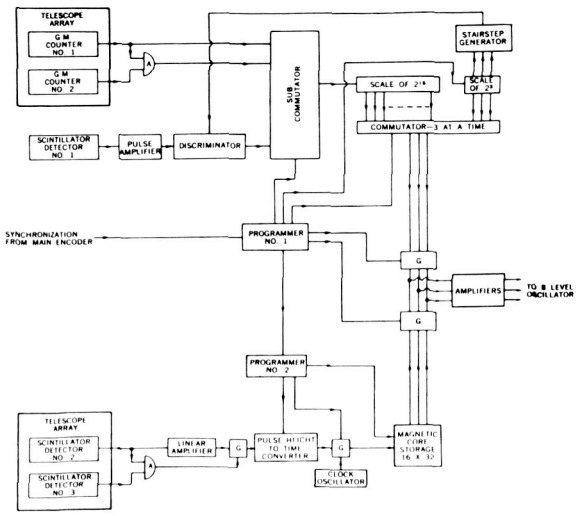


FIGURE 7.—Block diagram of the Explorer XII cosmic ray experiment.

GM counter coincidence events were accumulated for about 1.6 seconds. During the next 0.96 second the scaler was read into the telemetry system for transmission to the ground by the 3-bits-at-a-time commutator. Then the single GM counter output was accumulated and read during similar intervals. The output of the thin scintillator detector, with the discriminator set at the first level, was accumulated and telemetered during the next 1.6 and 0.96 second intervals respectively; and this was repeated for discriminator levels two through eight. Thus a complete set of counting rates for these two detectors was obtained in 25.6 seconds. This entire sequence was repeated 12 times in slightly more than 5 minutes (307.2 seconds exactly).

While these rates were being measured, pulses from the double scintillator telescope were being analyzed. The analyzer included a linear amplifier to increase the amplitudes of the PM tube pulses; a linear gate to block all pulses not resulting from an acceptable coincidence event; a pulse height-to-time converter from which a pulse train consisting of a number of pulses proportional to the initial pulse amplitude was obtained for each event; and a storage system. At the end of the five minute analysis and storage interval, the pulse height spectrum was contained in a 32 channel by 16 bit magnetic core storage matrix. The count capacity for each channel was, therefore, $2^{16}-1$ or 65,535 counts.

Following the accumulation of these data the "analyze and store" operation was terminated, the accumulation of counts from the GM counters and thin scintillator was interrupted, and the contents of the storage matrix were telemetered two bits at a time. The readout was repeated a second time; and this time the memory was erased when read. The first half of the erased memory was telemetered a third time so that the completeness of the erasure could be checked. The complete readout ($2\frac{1}{2}$ times) required 102 seconds. Following the readout, the system reverted to the "analyze and store" mode, and the data system again telemetered the counting rates from the GM counters and thin crystal scintillator. The entire cycle, including the obtaining of the counting rates of the GM counters, the integral spectrum from the thin scintillator, and the differential spectrum from the scintillator telescope, was repeated every 6.66 minutes.

The complete Explorer XII system described above, including the detectors, conditioning equipment, and pulse height analyzer, was fabricated in the form of five subassemblies weighing a total of 12.8 pounds, requiring about 1.4 watts of electrical power, and utilizing 530 transistors. The

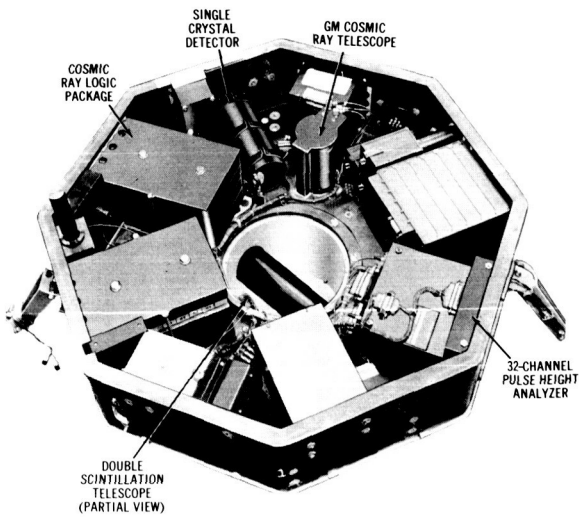


FIGURE 8.—The Explorer XII instrument tray. The units comprising the cosmic ray experiment are darkened. Not shown are the solar paddles, cover, and magnetometer boom and coil assembly.

subassemblies have been darkened in the photograph of the satellite instrument shelf (figure 8). During the satellite's operating life (113 days) this equipment obtained a large amount of interesting new data about the primary cosmic rays and solar particles.[2,3]

THE ECCENTRIC ORBITING GEOPHYSICAL OBSERVATORY (EGO) EXPERIMENT

An improved cosmic ray experiment is being prepared for launch aboard the first Eccentric Orbiting Geophysical Observatory (EGO) in mid-1963. This experiment will determine the relative abundances of nuclei of atomic numbers 1 through 8 (oxygen) and the energy spectrum for each, over the approximate energy range 11 to 90 Mev/nucleon. The experimental equipment employs scintillator crystals (figure 9) to measure independently the rate of energy loss and total energy of each particle. A thin CsI(Tl) crystal A of 0.45 gm/cm^2 thickness (1 mm) measures a ΔE which is approximately proportional to the rate of energy loss. The particles which stop in the thick (9 gm/cm^2) scintillator B produce an output that is nearly proportional to $E-\Delta E$, where E is the initial energy of the particle. A guard scintillator C of thickness 0.8 gm/cm^2 surrounds the thick scintillator, and a coincidence circuit establishes the condition that each incident particle produce outputs from both the ΔE and $(E-\Delta E)$ scintillators, but not from the guard scintillator B, in order that it be analyzed. This establishes the geometry for the detector and ensures that energetic particles passing completely through the thick scintillator are not analyzed.

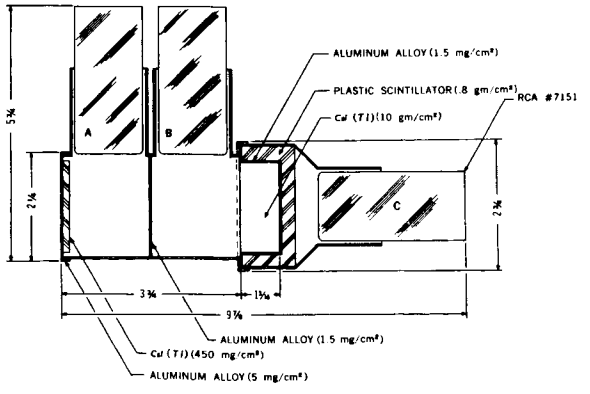


FIGURE 9.—The EGO nuclear abundance detector configuration. The light from the two CsI crystals is diffusely reflected by internal white surfaces onto the photocathodes of the upper two PM tubes.

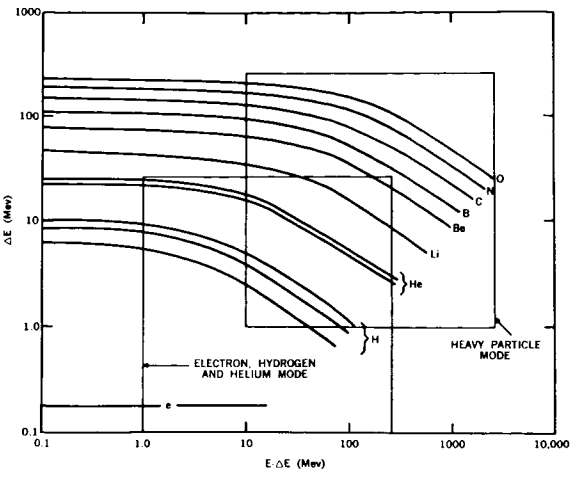


FIGURE 10.—The energy loss in the thin ΔE crystal as a function of energy loss in the thick $(E-\Delta E)$ crystal. The two squares indicate the regions to be covered by the EGO experiment.

The manner in which this detector determines the energy spectra of the individual nuclear constituents can be seen with the aid of figure 10, in which the energy lost in the ΔE scintillator is plotted as a function of the energy lost in the $(E-\Delta E)$ scintillator for the various nuclei. In practice the measurements are scattered about the curves shown—this is due to Landau scattering, the variation in path length through the thin scintillator for different angles of incidence, the variation in light transmission between the crystals and the photocathodes of the PM tubes, and the uncertainties of the measurements of the PM tube pulse heights. These effects combine to produce approximately a 15 percent resolution, measured at the half height of a 0.5 Mev line. This resolution is adequate to separate electrons, protons, deuterons, tritons, alpha particles and lithium, beryllium, boron, carbon, nitrogen, and oxygen nuclei. The separation of the He_3 and He_4 lines is marginal, but may be possible if the two are approximately equal.

The two squares in figure 10 indicate the areas in the ΔE vs $(E-\Delta E)$ field covered by the EGO experiment. The lower energy region represents

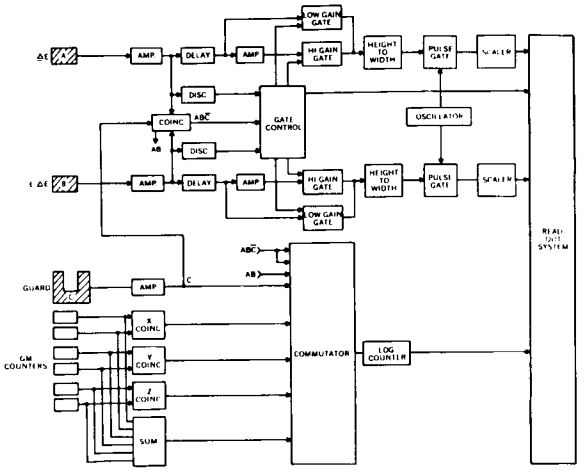


FIGURE 11.—Block diagram of the EGO nuclear abundance experiment.

the ranges normally covered by the pulse height analyzers. However, if a pulse from either the ΔE or the $(E-\Delta E)$ crystal exceeds the upper bounds of this region, the gains of the amplifiers in the analyzers are automatically reduced. The field covered by this configuration is represented by the upper energy, or heavy nucleon, square. The terminations of the high energy ends of the curves results from the action of the guard scintillator. The energies at these termination points are those required for the particles to pass completely through the $(E-\Delta E)$ scintillator but not enter the guard scintillator.

In this experiment a rather high resolution analysis of the pulse heights from the ΔE and $(E-\Delta E)$ photomultiplier tubes is necessary. The system block diagram (figure 11) shows the operation of the three dimensional (two parameter) 256 channel analyzer developed for this task. Scintillator pulses A, B, and C and a signal from a busy bistable multivibrator in the gate control circuit produce a gating signal at the coincidence circuit output if the A and B pulses are above the coincidence circuit threshold, the C pulse is below the threshold, and the busy bistable is in the "not busy" state. This coincidence pulse is then steered by the gain setting circuit: Normally the coincidence pulse passes to the pair of high gain linear gates; but if either the A or B pulse amplitudes exceed a threshold value, then the coincidence pulse is routed to the two low linear gates. A marker bit is telemetered with the data to indicate the gain setting for each event analyzed.

The A and B pulses are amplified, shaped, and delayed. The amplifiers are highly stable charge-input circuits employing very large amounts of feedback. Delay line shapers are employed to produce standard 2 μsec wide, squared pulses in order to remove the effects of changes in PM tube pulse characteristics. Additional pulse amplifiers are used in the high gain setting. The gate outputs charge the capacitors in the sweep circuits of the height-to-width converters. Sweep threshold detectors produce clock gating pulses, which permit numbers of clock pulses proportional to the amplitudes of the A and B pulses to emerge from the analyzers. Careful synchronization of the clock, busy bistable, and sweep circuit prevents jitter of the output pulse numbers. The busy bistable is set "busy" at the instant of occurrence of the clock pulse immediately following the charging of the sweep capacitor. At the same time, the clock pulse output gate is opened. The discharging of the sweep capacitor through a constant current generator is held off until this time; thus, the sweep always begins coincidently with the beginning of the first output clock pulse. The clock pulse output gate is closed when the sweep voltage reaches a threshold value.

The pulses from the clock pulse output gates are counted by two binary scalers. Eight stages are used, giving 256 channel pulse height resolution. Following the analysis of an acceptable event, binary numbers proportional to the A and B pulses reside in these scalers, and the busy bistable is in the busy state, holding off the analysis of additional events.

The data handling system in the EGO spacecraft is a PCM time-sharing system. Two main multipliers or commutators, located in two data system equipment groups (EG) are used. Each has 128 data inputs, corresponding to 128 nine-bit words in the telemetry frame. The words are entered into the equipment group designated by the signal on a switch signal line—the absence of a signal on this line means "feed EG 1" while the presence of a signal means "feed EG 2".

A number of lines connect the experiment to the data handling system. A data line is required for each equipment group to carry the serial

binary data from the experiment to the data system. A shift pulse line from each equipment group causes the bits at the experiment output to change in synchronism with the data system. Inhibit (or gating) signals from each telemetry word route the data from the experiment to the proper positions in the telemetry format.

For this experiment the data are entered into the data handling system by a shift register. When the data system is ready to begin accepting data, the contents of one of the two binary scalers are parallel-transferred into the shift register, and the scaler is reset. The contents of the shift register are then shifted serially into the data system in synchronism with the shift pulses. Then the contents of the other scaler are transferred and shifted. The contents of the two scalers are entered into adjacent words in the telemetry frame. Following the completion of this entry, the busy bistable circuit is reset to "not busy" and the system is free to accept another event for analysis.

Several auxiliary indicator bits are included in the data. The gain setting is entered into the ninth bit of the A binary scaler and, whenever either binary scaler overflows because an event occurs above or to the right of the low gain square of figure 10, a "data error" bit is entered into the ninth bit of the B binary scaler. In addition, a data error is indicated if the data system attempts to read the binary scalers while an analysis is still in process. To prevent the beginning of an analysis while the data system is still reading the experiment, the busy bistable is set "busy" at the beginning of the readout if this has not already been done by the occurrence of an acceptable event in the analysis interval preceding the readout.

The EGO system will be capable of accepting and analyzing an event every second when the data are being recorded by the spacecraft data storage system. Operation at 8 or 64 events per second will occur whenever the data are transmitted directly to the ground receivers without on-board storage. The dead time of the analyzer itself is approximately $10+2N$ microseconds, where N is the higher of the two numbers produced in the binary scalers. Of course, the system is also dead whenver the busy bistable is "busy."

A secondary system is included in this experiment to provide additional information about the nuclear abundance measurements, and to provide an independent measurement of the flux of relatively high-energy particles by a simple monitoring detector. An eight position commutator sequentially connects eight sources to a scaler. The pulses from the first source are counted for a known period and the contents of the scaler are read into the data system. Then the pulses from the second source are counted and read, etc. The eight inputs are as follows: (1) The input is left open, to assist in establishing frame synchronization. (2) The C guard detector pulses are counted to allow measurement of high fluxes of charged particles which might effect the accuracy of the primary measurements, and to give a measurement of the omnidirectional intensity. (3) The coincidence circuit ABC pulse rate is measured to determine the number of events which would have been analyzed if the instrument had a zero dead time. (4) The AB coincidence events are counted to give a rough measurement of the directional intensity. (5), (6), (7) The X-, Y-, and Z-axes GM counter telescope rates, respectively, are measured. (8) The sum rate from the six GM counters is measured to give a rough measurement of the omnidirectional flux.

The GM counter array forms a simple monitoring instrument to check the validity of the primary detector's measurements, to allow direct comparisons of results obtained on other flights on which similar arrays are flown, to provide measurements of the cosmic ray intensity, and to determine the directional characteristics of the cosmic rays. Each axis of this array is similar to the Explorer XII GM telescope described above and shown in figure 6.

The scaler which counts these pulses has a rather unique quasi-logarithmic characteristic which provides a large dynamic range while retaining a fixed accuracy, with the use of only nine binary data bits. The word obtained from the scaler has two parts: a number consisting of N binary bits and an exponent consisting of E binary bits. The count capacity $n_{\max}$ of the circuit is, in general

$$n_{\max}=2^N(2^{2^E}-1).$$

In this case $N=5$ and $E=4$, so that $n_{max}=2^{21}-2^5=2{,}097{,}120$. This scaler is read by the spacecraft data system in the same manner as the analyzer scalers described earlier.

This complete experiment is packaged in two assemblies, a cubical main package eight inches on a side, and a GM telescope package approximately 4×4×5.5 inches. The total weight of the experiment is 10.5 pounds, and it requires 1.7 watts of power. The basic detector assembly has been flown on balloon flights several times during the past year to check its characteristics.[4] These flights confirm the derivation of the detector parameters, and indicate that high resolution analysis of the charge spectrum will be possible.

CONCLUDING REMARKS

As more becomes known about a particular space phenomenon, experiments to study the phenomenon tend to assume a more complex character so that increasingly detailed information can be obtained. Because the cosmic rays have been investigated for many years, a great deal is already known about them. The experiments described in this paper were designed to increase the amount of detailed information about the energy and charge spectra of low energy cosmic rays. The Explorer XII instrument represents the first use of a medium-resolution differential analyzer and core memory system in space. The EGO system is an extension of this earlier work to make possible two parameter analysis with much higher resolution. It can be expected that instruments of this type will have wide application in space research as it becomes possible to make higher quality measurements of other phenomena.

ACKNOWLEDGMENTS

The authors are indebted to Mr. R. Schumann of Nuclear Data, Inc., for assistance in the early phases of the development of the two-parameter pulse height analyzer system, and to Dr. D. A. Bryant, who has been active in the development of the EGO detector array.

REFERENCES

1. Desai, U. D., Van Allen, R. L., and Porreca, G., "Explorer XII Rays," *Proc.* 1962 *National Telemetering Conf.*, Vol. I: Session 7, Article 5, May 1962, pp. 1–15.
2. Bryant, D. A., Cline, T. L., Desai, U. D., and McDonald, F. B., "Cosmic Ray Observations in Space," To be published in Proc. Third Internat. Space Sci. Symp., May 1962.
3. ———, ———, ———, and ———, "Explorer XII Observations of Solar Cosmic Rays and Energetic Storm Particles Following the Solar Flare of 28 September 1961," *Geophys. Res.* (in publication, 1962).
4. ———, Ludwig, G. H., and McDonald, F. B., "A Scintillation Counter Telescope for Charge and Mass Identification of Primary Cosmic Rays," *IRE Trans. on Nuclear Science*, NS–**9**(3): June 1962.

ON THE ZENITHAL DISTRIBUTION OF EXTREMELY HIGH ENERGY COSMIC RAY MUONS IN THE ATMOSPHERE

KAICHI MAEDA

Goddard Space Flight Center

The approximate calculation of high energy cosmic ray muon intensities at sea level which is given by Smith and Duller is improved by a rigorous calculation for larger zenith angles and higher energies, starting from the integral formula instead of the so-called diffusion equation of cosmic ray mesons in the atmosphere. Similar computations are done to show the contribution of kaons; the results are compared with those of Zatseptin and Kuz'min who used a different attenuation mean free path. The largest difference between the contributions of pions and kaons appears for muon energies of 100–1000 Gev, as pointed out by Zatseptin and Kuz'min. It is shown that measurements of obliquely incident high energy muons are helpful in determining the energy dependence of the attenuation mean free path of charged mesons in the atmosphere rather than to distinguish the muon parent particles. To resolve the arguments regarding the additional source of high energy cosmic ray muons, it is proposed that these observations be extended to mountain altitudes.

1. INTRODUCTION

As pointed out by several authors (Smith and Duller, 1959; Zatseptin and Kuz'min, 1961; Sheldon and Duller, 1962) the zenith angle distribution of high energy cosmic ray muons is quite different from that of low energy ones. The direction of maximum intensity at sea level shifts toward the horizontal direction with increasing muon energy from the verticle direction at the low energy.

The change of zenith angle dependence was calculated by Smith and Duller from 40 to 160 Gev of muon energy with respect to a flat atmosphere. Obviously this method fails at large zenith angles. Zatseptin and Kuz'min (1961) extended calculations to 10^5 Gev, taking into account the curvature of the earth's atmosphere. They calculated the contribution of muons produced not only from pion decay, but also from kaon decay, the importance of which has been already discussed by many authors (Smith and Duller, 1959; Duthie et al., 1960; Rodgers, 1961).

In the calculation by Zatseptin and Kuz'min, the attenuation mean free path of cosmic ray pions and kaons is assumed to be the same as that of nucleons, 75 g/cm^2. However, the attenuation mean free paths, occasionally called nuclear absorption mean free paths, of pions and of kaons in air are not known accurately, particularly at such high energies above the 100 Gev region.

On the other hand, calculations of the intensity of cosmic rays in the atmosphere have been based essentially on the so-called diffusion equation for these unstable particles in the atmosphere. As can be seen in the following sections, the solutions of these differential equations are not accurate enough to describe the diffusion of the cosmic ray particles with large zenith angle incidence in the atmosphere, because the local zenith angle changes with atmospheric depth (Fig. 1).

In this paper, therefore, calculations are to be done as follows:

(i) Instead of diffusion equations for pions and kaons, integral expressions will be used from the beginning, revising the author's previous computation (Maeda, 1960*). By starting with

*Later in this paper, this will be referred to as M.1.

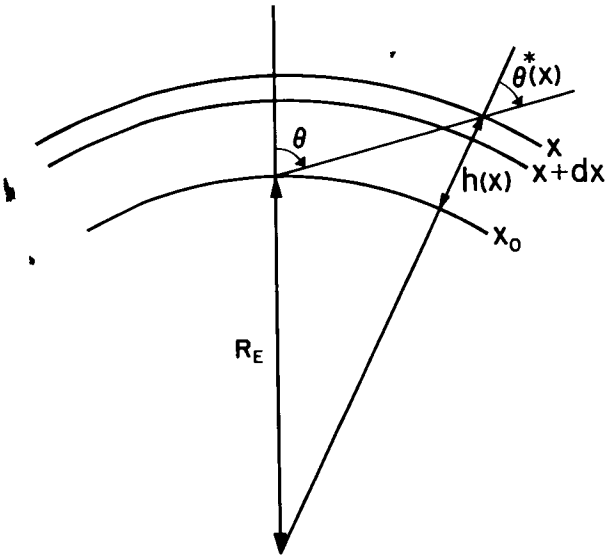


FIGURE 1.—Geometrical relation between the zenith angle θ at the depth x_0 and the local zenith angle $\theta^*(x)$ at the depth x in the atmosphere. R_E is the earth's radius and $h(x)$ is the height of the level of atmospheric depth x.

the integral form, the effect of the curvature of the earth's atmosphere as well as the non-linear energy loss of high energy particles and change of atmospheric density can be rigorously taken into account. It can be shown that these integral forms become solutions of diffusion equations which are used by the above-mentioned authors under particular approximations.

(ii) By calculating two cases for the attenuation mean free path of the pion, the effect of this unknown quantity on the zenith angle distribution will be shown as a parameter, which can be determined experimentally, although qualitative discussions are given by Smith and Duller.

(iii) The zenith angle distribution of cosmic ray intensity changes with atmospheric depth. Therefore, the calculations are extended to mountain altitudes to see if any significant change from the zenithal distributions at sea level appears at these heights.

(iv) To see the difference between the contributions of pions and kaons, calculations for both are done by using the same attenuation mean free path value of 120 g/cm^2.

It should be noted that the increase of high energy muon intensity in an oblique direction is simply due to the increase of the path lengths of pions (and of kaons) within the upper atmosphere, where the rate of decay into muons is larger than in the lower atmosphere because of low density of the air. At low energy, the muon intensity decreases because of decay along a prolonged path in a horizontal direction. At intermediate energy, a maximum intensity occurs at a certain oblique direction given by these two competing effects.

2. DIFFERENTIAL INTENSITIES OF COSMIC RAY MESONS IN THE EARTH'S ATMOSPHERE

An important kaon mode for muon production in the atmosphere is the K-mu-2 mode ($K_{\mu 2}$) for which decay is the same type as that of charged pions; its branching ratio is known to some extent (Ashton and Wolfendale, 1963). Therefore, in this paper only this mode will be considered; then the mathematical expressions are the same for pions and kaons, except for the constant value such as rest mass and mean lifetime of these mesons.

In the following, therefore, suffix i will mean these cosmic ray mesons which are parent particles of muons where $i=\pi$ and $i=k$ stand for pions and kaons, respectively. Since scattering processes are negligible for these high energy particles, the differential intensity of cosmic ray mesons $n_i(E_i, x, \theta^*)$ between E_i and $+dE_i$ at the atmospheric depth x in dx with zenith angle θ^* in $d\theta^*$ is given by

$$n_i(E_i, x, \theta^*) = \int_0^x \exp\left[-\int_{x'}^x \left(\frac{B_i}{E_i'' x''} + \frac{1}{L_i} \right) \frac{dx''}{\cos\theta^*} - \int_0^{x'} \frac{dx''}{L_n \cos\theta^*} \right] \frac{F(E_i'')}{L_c \cos\theta^*(x')} dx' \quad (2.1)$$

where $F(E_i'')$ is the differential production spectrum of cosmic ray mesons, E_i'' is the energy at the depth x'', which becomes E_i at the depth x, (relations between $E_i''(x'')$ and $E_i(x)$ are given in Appendix I under several types of approximation); L_c, L_n and L_i are the collision and attenuation mean free paths of primary cosmic ray particles (nucleons) and the nuclear absorption

mean free path of mesons in the air, respectively, and θ^* is the local zenith angle at the atmospheric depth x, i.e.,

$$\cos\theta^* = \left[1 - \frac{\sin^2\theta}{\{1+[h(x)/R_E]\}^2}\right]^{1/2} \quad (2.2)$$

with $h(x)$; the height of level x above the earth's surface and R_E; the earth's radius, 6370 km (figure 1). The decay factor, B_i, is given by

$$B_i = (m_i c/\tau_i)\cdot[x''/\rho(x'')] \quad (2.3)$$

where $\rho(x'')$ is air density at the depth x'', m_i is the rest mass of a cosmic ray meson, i.e., $m_\pi = 140$ Mev/c^2 and $m_k = 494$ Mev/c^2, τ_i is the mean life of a charged cosmic ray meson, i.e., $\tau_\pi = 2.55\times10^{-8}$ sec and $\tau_k = 1.22\times10^{-8}$ sec.

In the case of isothermal atmosphere, which is good approximation for the stratosphere, where most of the decay of cosmic rays into muons takes place, the factor $x''/\rho(x'')$ is constant. Then, it can be assumed that

$$B_\pi = 117 \text{ Gev} \quad \text{and} \quad B_k = 858 \text{ Gev}. \quad (2.4)$$

It should be noted that the term

$$F(E_i'')\, dx'/L_c \cos\theta^*(x')$$

is the production rate of the mesons of energy E_i'' at the depth x' within a layer of thickness dx' by the unit intensity of primary cosmic ray with incident zenith angle θ^* at the depth of x' in the atmosphere. The first exponential term of the integrand indicates the decrease of meson intensity in the atmosphere due to decay and nuclear absorption in the air, and the second exponential term corresponds to the decrease of meson-producing particles with an increase of atmospheric depth.

If the change of the zenith angle of the path of cosmic ray particles with atmospheric depth is neglected, i.e., if $\theta^* = \theta =$ constant is assumed, then the integral of Eq. (2.1) has the somewhat simple form which is given for a differential intensity of cosmic ray mesons diffusing in a *flat* atmosphere. These are solutions of the cosmic ray diffusion equations with a certain approximation as shown in Appendix I.

3. DIFFERENTIAL INTENSITIES OF OBLIQUELY INCIDENT COSMIC RAY MUONS

3.1. Energy Relation Between Muons and Parent Mesons

Provided that the total energy and momentum of mesons in laboratory system are given by E_i and P_i, respectively, then the total energy of muons in this system, E_μ, is given by Lorentz formula

$$E_\mu = E_{\mu 0}(E_i/m_i c^2) + cp_{\mu 0}\cdot(p_i/m_i c)\cos\theta_0 \quad (3.1.1)$$

where θ_0 is the angle of emission of muons in the rest system of the parent meson measured from the direction of the momentum of the meson. $E_{\mu 0}$ and $p_{\mu 0}$ are the total energy and momentum of muons in the rest system of the parent meson, given as follows (Ascoli, 1950);

$$E_{\mu 0} = \tfrac{1}{2}(m_\mu c^2)[(m_i/m_\mu) + (m_\mu/m_i)] \quad (3.1.2)$$

and

$$p_{\mu 0} = \tfrac{1}{2}(m_\mu c)[(m_i/m_\mu - (m_\mu/m_i)] \quad (3.1.3)$$

where m_μ is the rest mass of the muon, 106 Mev/c^2.

By making use of the so-called extreme relativistic approximation, i.e., $p_i c = E_i$, which causes no significant errors in the present calculation, it can be seen that the range of the energy E_i of parent mesons for a given energy of muons E_μ extends from E_μ to $(m_i/m_\mu)^2 E_\mu$.

Emission of muons in the rest system of mesons can be assumed to be isotropic; then the probability $P(E_\mu)dE_\mu$ that the energy of muons is between E_μ and $E_\mu + dE_\mu$, is obtained from Eq. (3.1.1),

$$P(E_\mu)dE_\mu = \tfrac{1}{2}\sin\theta_0\cdot d\theta_0 = (m_i/2p_{\mu 0}p_i)\, dE_\mu$$

By using Eq. (3.1.3), this can be written as

$$P(E_\mu)\, dE_\mu = [1-(m_\mu/m_i)^2]^{-1}(dE_\mu/E_i) \quad (3.1.4)$$

Therefore, the number of muons $n_\mu(E_\mu, x, \theta^*)$, produced in the layer between x and $x+dx$, with energy E_μ in dE_μ and in the direction θ^* in $d\theta^*$, is

$$n_\mu(E_\mu, x, \theta^*) = \int_{E_\mu}^{E_\mu r_i^2} \frac{B_i \sec\theta^*(x)}{E_i\cdot x} P(E_\mu) \cdot n_i(E_i, x, \theta^*)\, dE_i \quad (3.1.5)$$

where $r_i = m_i/m_\mu$,

It should be noted that the decay factor, $B_i/E_i x$, which is omitted in Ascoli's paper (1950), is needed to explain the fact that the energy spectrum of underground muons changes from $E^{-1.8}$ to $E^{-2.8}$ below about 700 m.w.e. (Hayakawa and Tomonaga, 1949).

The integration of the right hand side of Eq. (3.1.5) can be approximated by the following simple form (Maeda and Wada, 1954)

$$n_\mu(E_\mu, x, \theta^*) \simeq [B_i \sec\theta^*(x)/E_\mu \cdot x] n_i(E_i', x, \theta^*) \tag{3.1.6}$$

where $E_i' = rE_\mu$. The accuracy of this approximation decreases with the power of the energy spectrum and the error is too large to use for the case of Kaon-produced muons as shown in Appendix II.

3.2. Rigorous Calculation of the Differential Intensity of Muons

The differential energy spectrum of muons $i_\mu(E, x_0, \theta)$ with energy between E and $E+dE$ at the atmospheric depth x_0 with zenith angle between θ and $\theta+d\theta$ is given by

$$i_\mu(E, x_0, \theta) = \int_0^{x_0} n_\mu(E_\mu, x, \theta^*) W(E, x_0, x, \theta)\, dx \tag{3.2.1}$$

where $W(E, x_0, x, \theta)$ is called the survival probability of muons in the atmosphere (Sands, 1950) which is the probability that the muons produced at the atmospheric depth x can arrive at the depth x_0 along the direction of zenith angle θ with energy E:

$$W(E, x_0, x, \theta) = \exp\left(- \frac{m_\mu c^2}{c\tau_\mu} \int_x^{x_0} \frac{\sec\theta^*(x')\, dx'}{E_\mu(E, x', \theta)\rho(x')} \right) \tag{3.2.2}$$

where τ_μ is the mean life of muons, 2.2×10^{-6} sec, $\rho(x)$ is air density at the depth x, and, as can be seen from Fig. 1, $\sec\theta^*$ is given by Eq. (2.2).

In the standard atmosphere, $\rho(x)$ and $h(x)$ are simple functions of x, as shown in W.1 $W(E, x_0, x, \theta)$ is shown in W.1 for several values of $\theta(0°-89°)$ and of $E(<43$ Gev). Survival probabilities of muons which are underground are shown by Hayman et al. (1962) for $E=200\sim$ 100,000 Gev.

Since the cosmic ray muons penetrate the dense lower atmosphere, the energy losses in the air as well as the change in local zenith angle along the path are no longer negligible for the oblique direction. Moreover, above several hundred Gev, the energy losses due to Bremsstrahlung and direct pair production cannot be neglected, as in the case of muon-producing parent mesons of such high energies. Expression of the energy loss rate of high energy muons including collision loss, radiation loss, and losses due to pair production and nuclear interaction as a function of muon energy has been given by many authors (Hayakawa and Tomonaga, 1949, Barrett et al 1952, Zatseptin and Kuz'min, 1961, Ashton 1961, Ozaki, 1962, Hayman et al 1962). The formula derived by Ashton (1961) seems most accurate. However, in the present calculation the formula (1.5) is used with the constants shown in Appendix I. Comparison of the energy loss rate formula given by several authors are shown in figure 9. From this figure one can see that the formula used in the present calculation is very similar to that of Zatseptin and Kuz'min (1961) and Ozaki (1962).

The expression and values of parameters of energy loss can be assumed to be the same as pions as given in Appendix I. Then, by using the similar relation given by Eq. (I.6) the energy at production, $E_\mu(E, x', \theta)$ is given as a function of x' for given values of E at x_0 and of θ at x_0 by

$$E_\mu(E, x', \theta) = [(a/b)+E] \exp\left(b \int_x^{x_0} \sec\theta^*(x'')\, dx'' \right) - (a/b)$$

$$\cong E + (a+bE) \int_x^{x_0} \sec\theta^*(x'')\, dx'' \tag{3.2.4}$$

By using this energy relation, the right hand side of Eq. (3.2.2) can be computed.

4. NUMERICAL CALCULATIONS AND RESULTS

If the production spectrum of mesons $F(E_i'')$ is known, the directional differential spectrum of muons $i_\mu(E, x_0, \theta)$ can be obtained by equations (3.2.1), (3.2.2), (3.1.5) and (2.1).

In Fig. 2 the differential energy spectra of muons at sea level are plotted, under the assump-

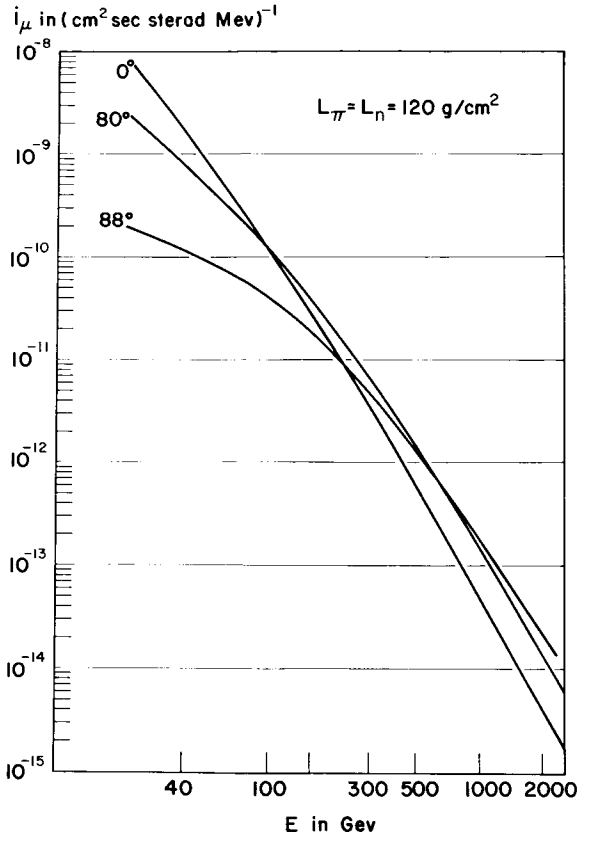


FIGURE 2(a).—The differential energy spectra of cosmic ray muons at sea level, normalized to the data of vertical differential intensity at sea level at 40 Gev. The attenuation mean free path of pions is assumed to be the same as that of nucleons, i.e., $L_\pi = L_n = 120$ g/cm^2.

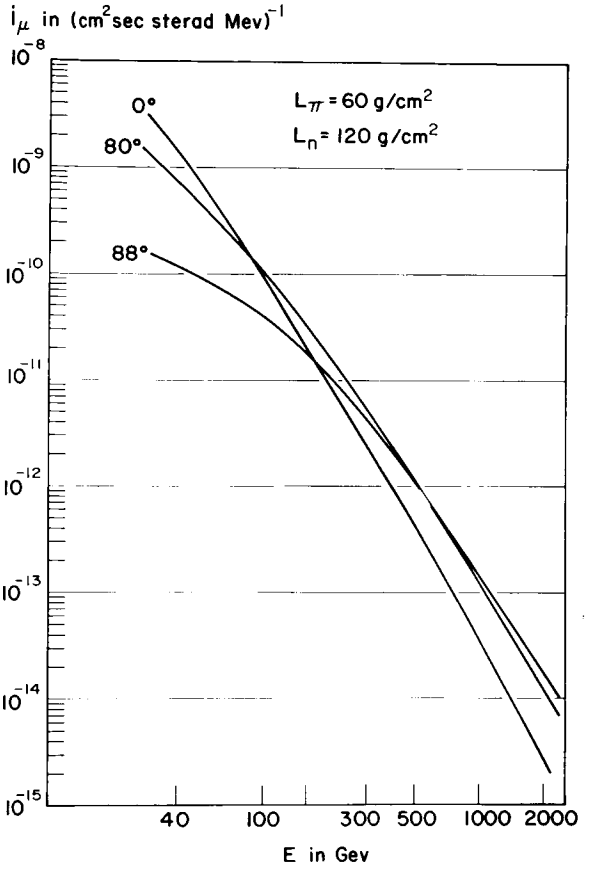


FIGURE 2(b).—The differential energy spectra of cosmic ray muons at sea level, normalized to the data of vertical differential intensity at sea level at 40 Gev. The attenuation mean free path of pions is assumed to be the same as that of nucleons, i.e., $L_\pi = 60$ g/cm^2.

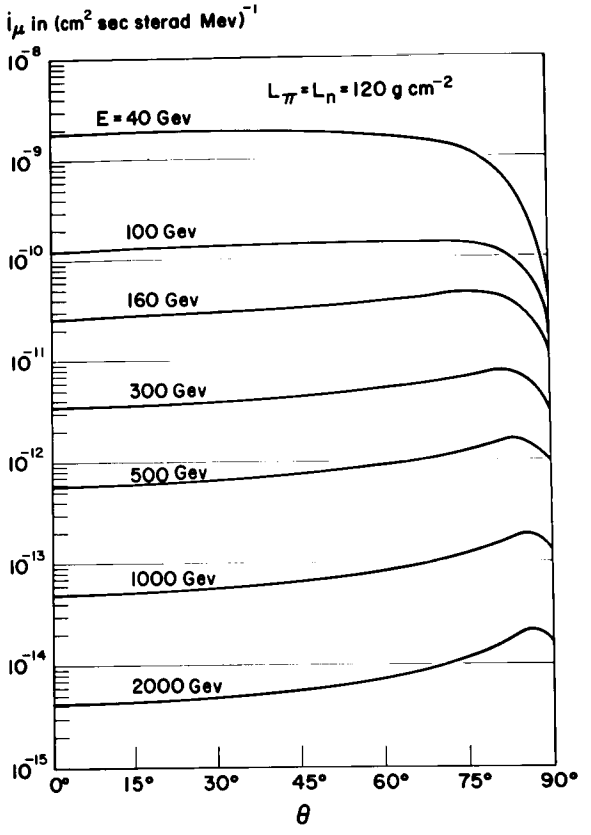


FIGURE 3(a).—The zenith angle dependence of differential intensity of muons, corresponding to figure 2(a).

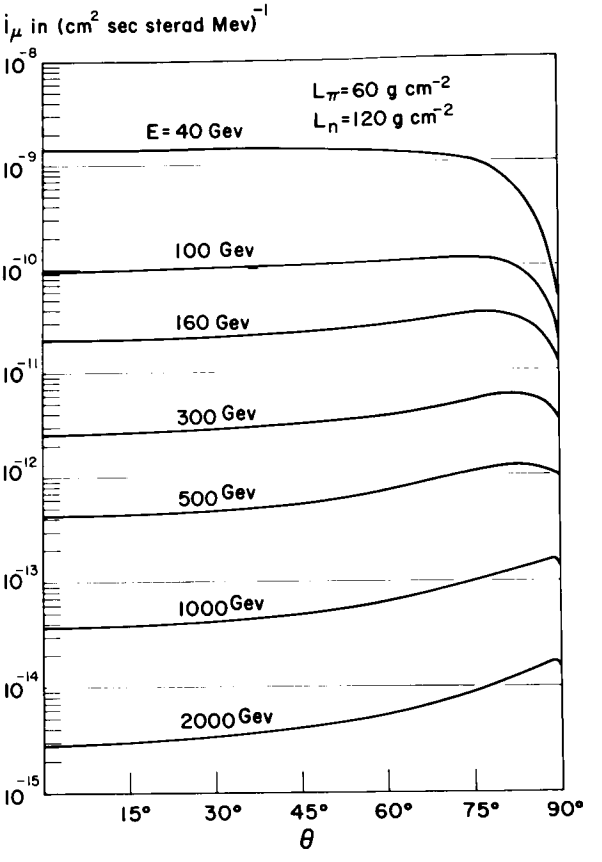


FIGURE 3(b).—The zenith angle dependence of differential intensity of muons, corresponding to figure 2(b).

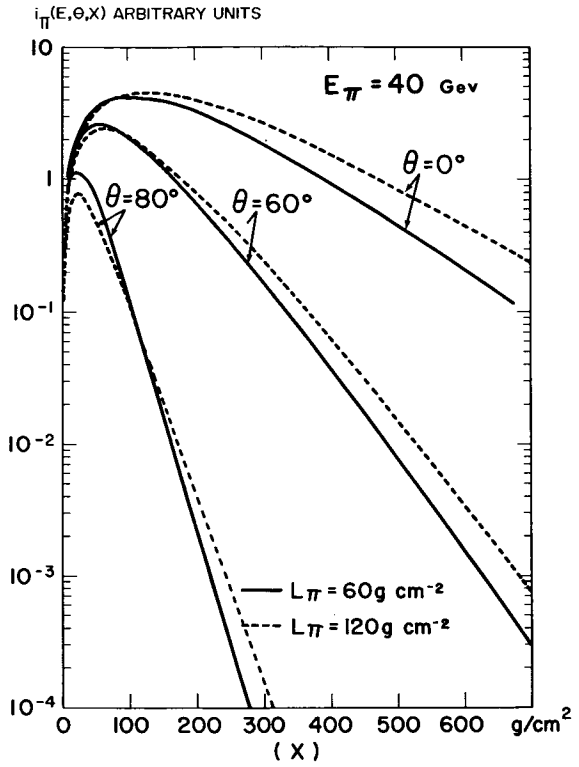


FIGURE 4(a).—The differential intensities of cosmic ray pions with an energy of 40 Gev in the upper atmosphere for different zenith angle, 0°, 60°, and 80°. Full lines and dashed lines correspond to L_π=60 g/cm² and L_π=120 g/cm², respectively.

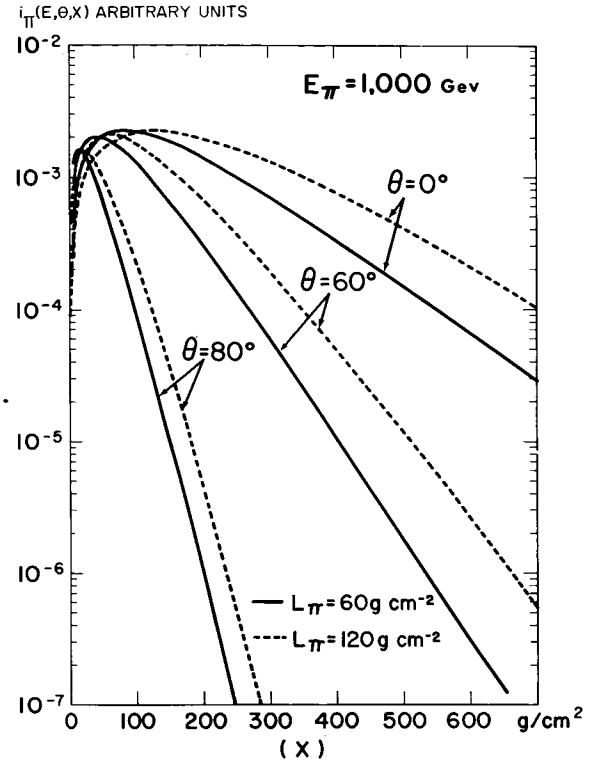


FIGURE 4(b).—The differential intensities of cosmic ray pions with an energy of 1000 Gev in the upper atmosphere for different zenith angle, 0°, 60°, and 80°. Full lines and dashed lines correspond to L_π=60 g/cm² and L_π=120 g/cm², respectively.

tion that all muons are produced by pion decay, and that the power of the production spectrum of pions is -2.7, i.e., $F(E'') = AE^{-2.7}$ where a constant A is left, arbitrarily, together with the collision mean free path L_c of primary particles in Eq. (2.1). The final results are normalized to the data of vertical differential intensity of sea level at 40 Gev (Woroney and Parry, 1954; Pak et al., 1961; Ashton et al, 1960). In Fig. 2(a), the attenuation mean free path of the pion is assumed to be the same as that of the nucleon, i.e., $L_\pi = L_n = 120$ g/cm^2, and in Fig. 2(b), it is assumed that $L_\pi = 60$ g/cm^2 and $L_n = 120$ g/cm^2. Fig. 3(a) and (b) are corresponding plots of the same results against the zenith angle. These results can be regarded as an extension of the calculations by Smith and Duller (1959) towards higher energy and larger zenith angle. Smith and Duller assumed the production spectrum of pions to be $AE^{-8/3}$.

To show the changes in the effective height of pion production due to the changes in incident zenith angle and attenuation mean free path, intensity versus depth curves calculated by Eq. (I.8) are plotted in Fig. 4(a) and (b) for $E = 40$ and 1000 Gev, respectively. From these figures it can be seen that the height of maximum intensity of pions shifts toward the higher latitude for large zenith angles, and that the decrease of the intensities of obliquely incident pions is less for high energy than for low energy pions. Since survival probability increases with energy, the effective height of muons which is the integrand of the right side of Eq. (3.2.1) shifts further upward with increasing zenith angle and energy (M.1 Fig. 16). Shift of the effective height of muon production with zenith angle is also shown by Zatseptin and Kuz'min (1961).

To magnify the change of muon intensity at large zenith angle, plots similar to those used by Zatseptin and Kuz'min, are given in figures 5 and 6. Namely, the relative differential intensities of muons $i_\mu(E, x_0, \theta)/i_\mu(E, x_0, 0)$ are plotted against $\cos\theta$, with energies of muons as parameter. In Fig. 5 the parameter is muon energies. (a) for $x_0 = 1030$ g/cm^2 (sea level), (b) $x_0 = 750$ g/cm^2 (mountain altitudes of about 2500 m above sea level), and (c) for $x_0 = 500$ g/cm^2 (around 5000 m above sea level). In Fig. 5, two cases of pion

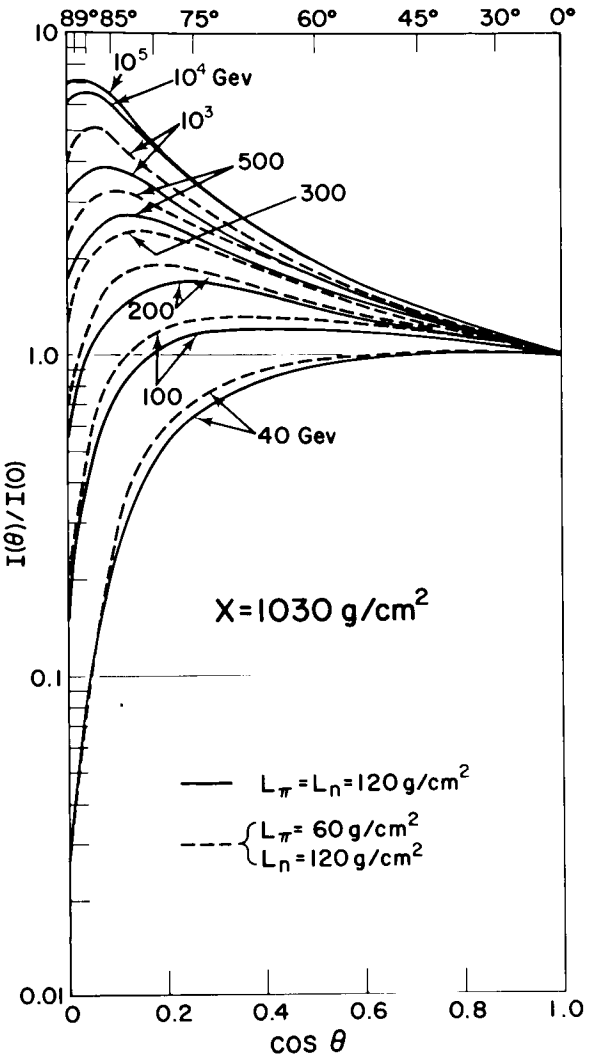


FIGURE 5(a).—Relative intensity of pion-produced muons normalized to vertical incidence vs. cos θ where θ is the zenith angle of incident direction at $x_0 = 1030$ g/cm^2. (Full lines and dashed lines correspond to $L_\pi = 120$ g/cm^2 and $L_\pi = 60$ g/cm^2, respectively.)

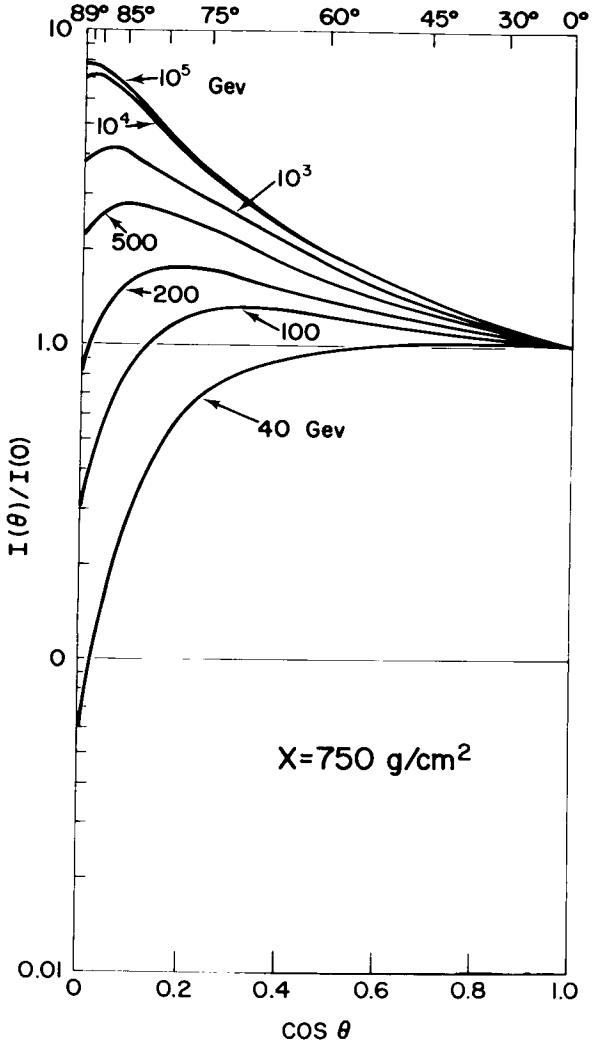


FIGURE 5(b).—Relative intensity of pion-produced muons normalized to vertical incidence vs. cos θ where θ is the zenith angle of incident direction at $x_0=750$ g/cm². (Full lines and dashed lines correspond to $L_\pi=120$ g/cm² and $L_\pi=60$ g/cm², respectively.)

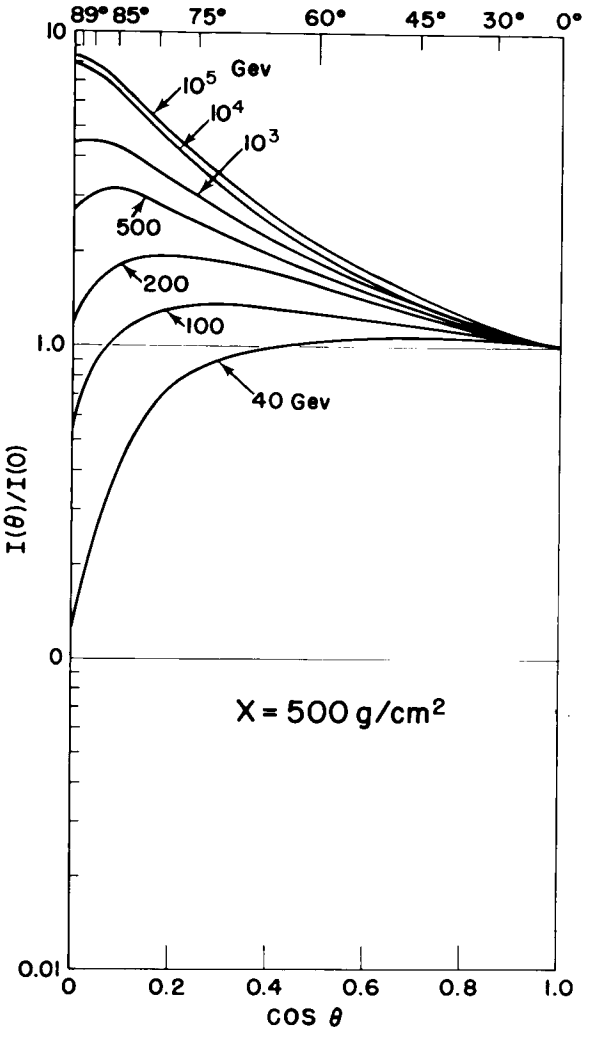


FIGURE 5(c).—Relative intensity of pion-produced muons normalized to vertical incidence vs. cos θ where θ is the zenith angle of incident direction at $x_0=500$ g/cm². (Full lines and dashed lines correspond to $L_\pi=120$ g/cm² and $L_\pi=60$ g/cm², respectively.)

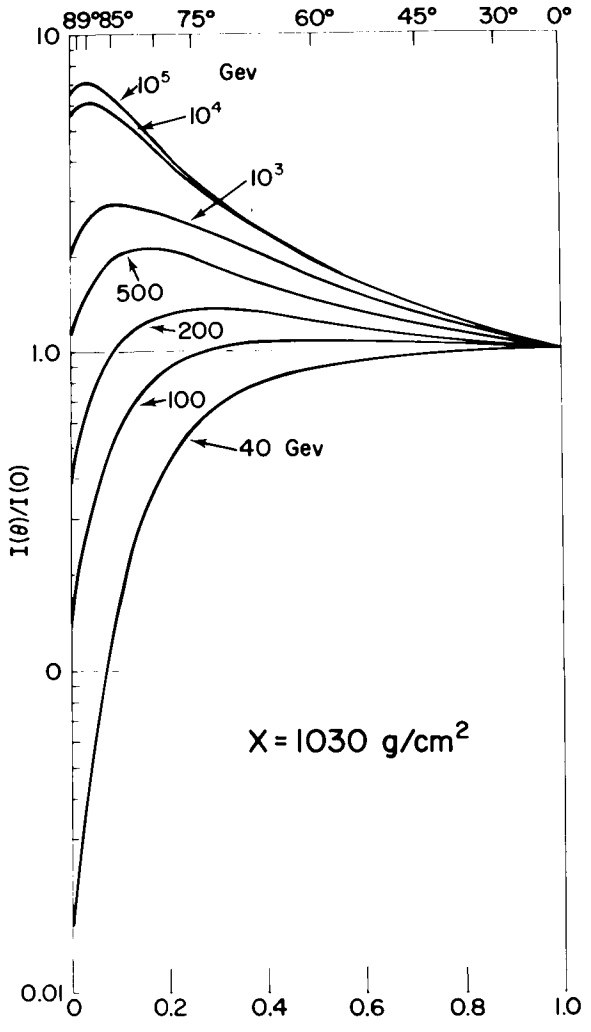


FIGURE 6(a).—Relative intensity of kaon-produced muons normalized to vertical incidence vs. cosθ at $x_0 = 1030$ g/cm^2.

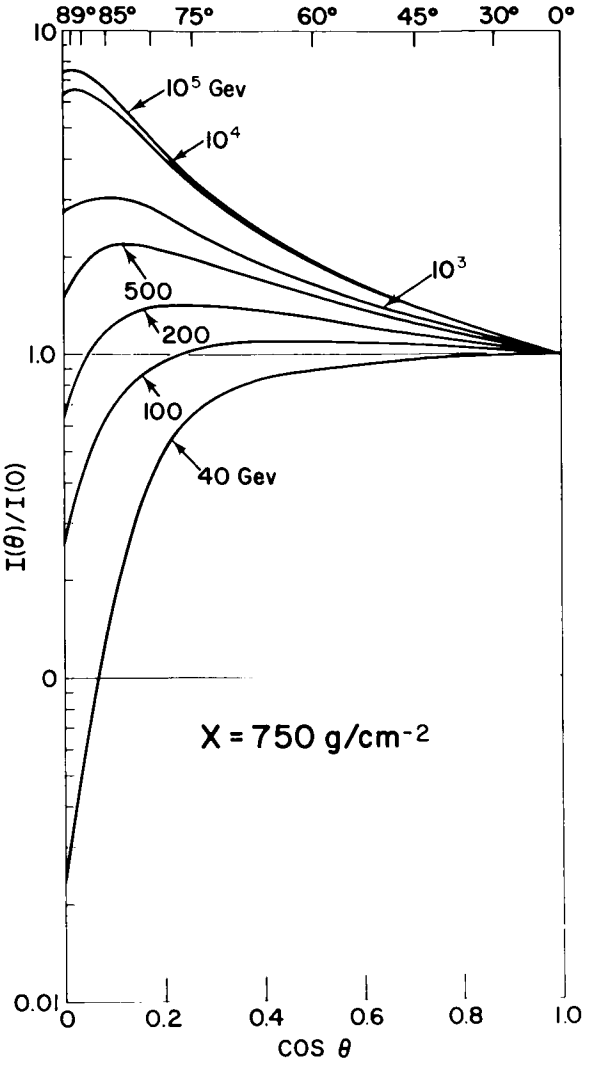


FIGURE 6(b).—Relative intensity of kaon-produced muons normalized to vertical incidence vs. cosθ at $x_0 = 750$ g/cm^2.

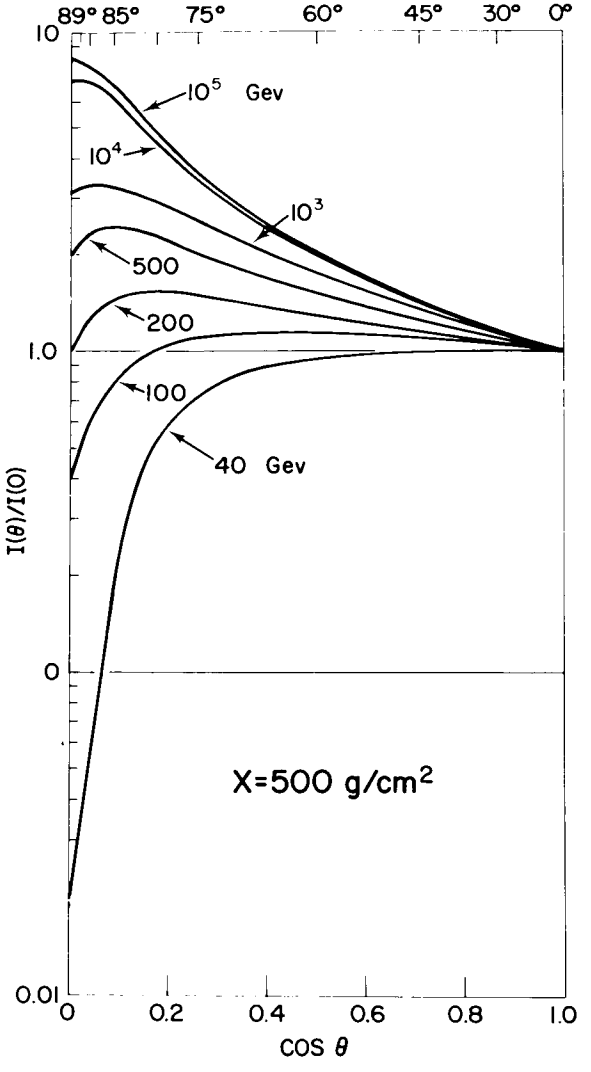


FIGURE 6(c).—Relative intensity of kaon-produced muons normalized to vertical incidence vs. $\cos\theta$ at $x_0=500$ g/cm².

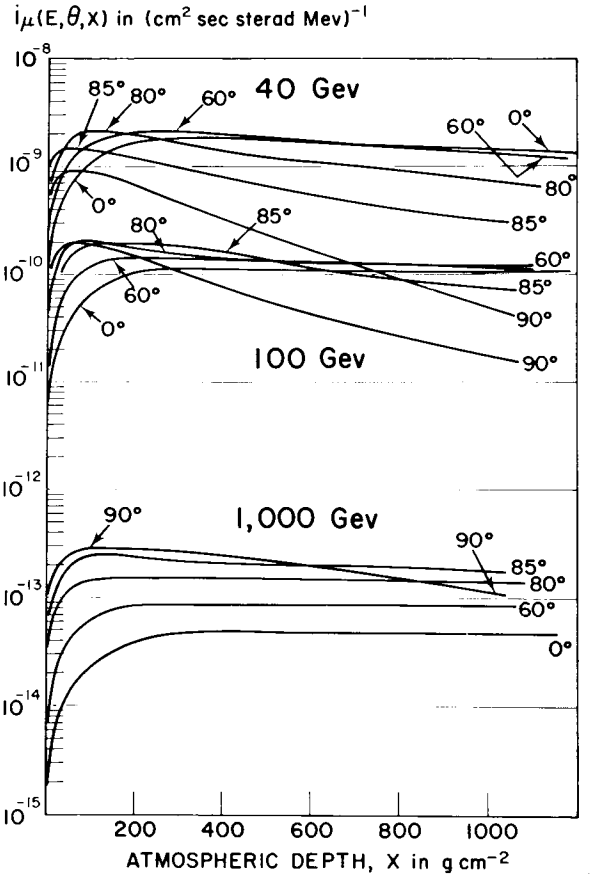


FIGURE 7.—Relative intensity of pion-produced muons vs. atmospheric depth x for $E_\mu=40$, 100, and 1000 Gev at sea level.

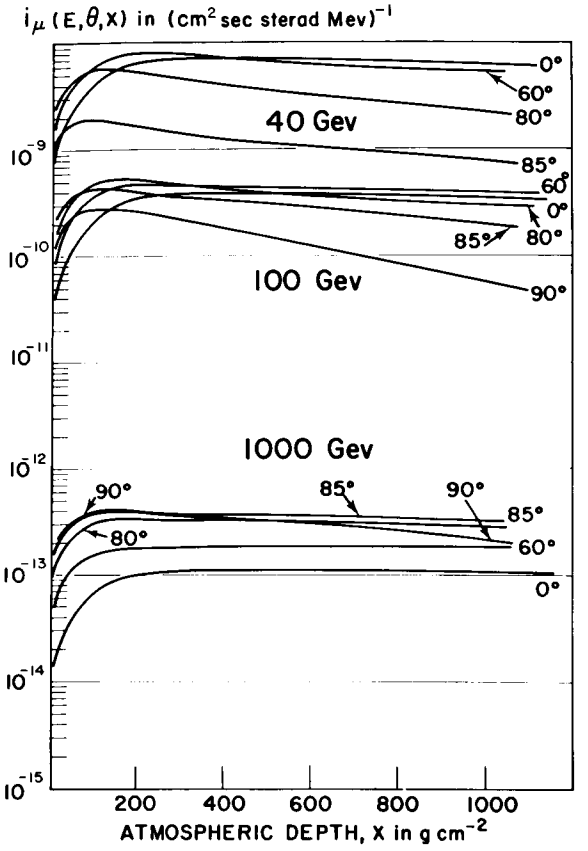


FIGURE 8.—Relative intensity of kaon-produced muons vs. atmospheric depth x (g/cm²) for $E\mu = 40$, 100, and 1000 Gev at sea level.

mean free path are shown for comparison. Figures 5(b) and (c) illustrate the trend of zenith angle dependence with decreasing atmospheric depth. Figures 6(a), (b), and (c) are similar figures, with the assumption that all muons are produced by kaon decay, the attenuation mean free path of the kaons being 120 g/cm². From these figures, it can be seen that the largest difference between pion-produced muons and kaon-produced muons is expected to be around 100–1000 Gev for muons in a large zenith angle, as pointed out by Zatseptin and Kuz'min. The relative intensity in the horizontal direction increases with altitude. To show this trend in high altitudes, the altitude dependences of pion-produced muons and kaon-produced muons for several different zenith angle are shown in Fig. 7 and 8 respectively. In Fig. 8, the production spectrum of kaons is assumed proportional to $E_k^{-2.95}$ (Ashton and Wolfendale, 1963). From Fig. 7 and 8, it can be seen also the height of the maximum intensity of muons shifts upward with increasing energy and zenith angle.

5. DISCUSSIONS

As mentioned before, the present calculation shows the correctness of the general statements on zenithal distribution of high energy muons given by Smith and Duller. We can summarize the present results as follows:

(i) The effect of decreasing L_i, the attenuation mean free path of muon-producing mesons, is qualitatively the same as the effect increasing the energy of muon-producing mesons on the zenith angle distributions of muons, i.e., the maxima shift to larger zenith angle and the intensities decrease, as discussed by Smith and Duller. By comparing Fig. 2(a) and (b), a quantitative difference of this effect can be seen for $L_\pi = 120$ g/cm² and 60 g/cm².

(ii) As pointed out by Zatseptin and Kuz'min, the largest difference between the contributions of pions and kaons on the zenith angle distribution of cosmic ray muons at sea level is found in the energy range from 100 to 1000 Gev. Above this energy range, the zenith angle distribution is simply dependent on the atmospheric structure.

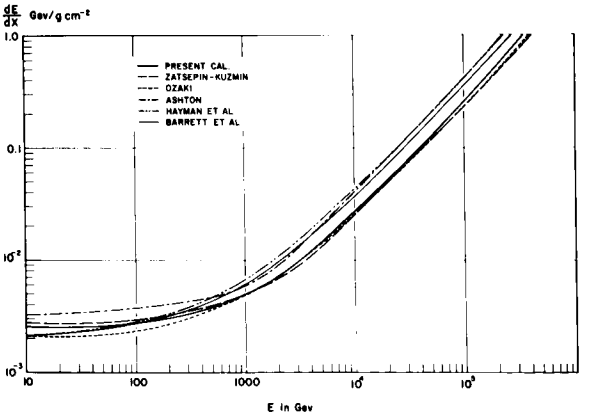


FIGURE 9.—Comparison of energy loss rate, dE/dx in Gev/g cm^{-2}, of muons vs. energy E with the one derived by several other authors. Full line corresponds to present calculation, while the others are as follows; – – – (Zatsepin-Kuz'min, 1961) ••• (Ozaki, 1962) —— (thin line, Barrett et al., 1952) —•—•— (Ashton, 1961) —••—••— (Hyman et al., 1962).

It should be noted that in this high energy region the effect of a density change in the atmosphere, which corresponds to a temperature change, is well established. An increase of temperature or a decrease of air density at each level in the atmosphere gives an increase of muon intensity due to the increase of the decay of muon-producing mesons (Maeda and Wada, 1954; M.1). On the other hand, the effect of pressure on the attenuation of muons is of course negligible in this high energy range (Wada, 1960).

(iii) The relative intensity of obliquely incident high energy muons increases with altitude. Since the attenuation mean free paths of cosmic ray mesons are not well known, observations of these high energy muons at mountain altitudes and comparison with sea level data will be very helpful in determining these quantities, particularly, as speculated by many authors (Rodgers, 1961; Duthie et al., 1961), the energy dependence of the attenuation mean free path of muon producing particles.

(iv) In the present calculation, the range-energy relation of muons is based on the simple formula (I.5), which gives nearly a value 80% less than that of Ashton's (1961), above 1,000 Gev as can be seen from Fig. 9. If Ashton's formula is correct, the energy of muons shown in the previous figures should be increased correspondingly. However, the relative shapes of zenith angle dependence shown in Fig. 5 and Fig. 6 remain the same, if the absolute value of muon-energy is shifted accordingly. The same consideration should be applied to the results given by Zatseptin and Kuz'min (1961), for which the energy loss rate is very close to the present formula and the one given by Ozaki (1962).

Appendix I

APPROXIMATIONS OF EQ. (2.1) AND CORRESPONDING DIFFUSION EQUATIONS FOR COSMIC RAY MUONS IN THE ATMOSPHERE

I-1. THE FIRST APPROXIMATION—NO ENERGY LOSS

In this case, $E_i'' = E_i =$ constant in Eq. (2.1) and

$$n_i(E_i, x, \theta^*) = \frac{F(E_i)}{L_c \cos\theta} \exp(-x/L_i \cos\theta) \int_0^x (x'/x)^{(B_i/E_i \cos\theta)} \exp[-(L_n^{-1} - L_i^{-1})(x'/\cos\theta)]\, dx' \quad \text{(I.1)}$$

It should be noted that this is a solution of the following differential equation with a boundary condition $n_i(E_i, 0, \theta^*) = 0$.

$$\frac{\partial n_i(E_i, x, \theta)}{\partial x} = \frac{F(E_i)}{L_c \cos\theta} \exp(-x/L_n \cos\theta) - \frac{n_i}{\cos\theta}\left(\frac{1}{L_i} + \frac{B_i}{E_i x}\right) \quad \text{(I.2)}$$

This is the same type of equation as used by Smith and Duller (1959) and Zatseptin and Kuz'min (1960) for the calculation of pion-intensity in the atmosphere.

I-2. THE SECOND APPROXIMATION—CONSTANT IONIZATION LOSS

In this case by the energy-range relation $E_i'' = E_i + a(x - x'') \sec\theta^*$, where a is the rate of ionization loss of pions in air. Then the integral (2.1) can be written

$$n_i(E_i, x, \theta^*) = \int_0^x \frac{F(E_i'')}{L_c \cos\theta} \exp(-x/L_i \cos\theta) \left(\frac{E_i}{E_i''} \cdot \frac{x'}{x}\right)^{(B_i \sec\theta)/(E_i + ax \sec\theta)} \exp[-(L_n^{-1} - L_i^{-1})(x'/\cos\theta)] \quad \text{(I.3)}$$

Under a boundary condition, $n_i(E_i, 0, \theta^*) = 0$, this is a solution of the following differential equation:

$$\frac{\partial n_i(E_i, x, \theta)}{\partial x} = \frac{F(E_i)}{L_c \cos\theta} \exp(-x/L_n \cos\theta) + \frac{a}{\cos\theta} \cdot \frac{\partial n_i(E_i, x, \theta)}{\partial E_i} - \frac{n_i}{\cos\theta}\left(\frac{1}{L_i} + \frac{B_i}{E_i x}\right) \quad \text{(I.4)}$$

---

this is essentially the same type of equation as used by Zatseptin and Kuz'min (1961) to describe muon diffusion in the atmosphere. Since a muon has no direct nuclear interaction, an attenuation term due to the nuclear absorption can be neglected completely, i.e.

It should be noted that the second term on the right side of Eq. (I.4), which corresponds to the change of differential intensity of cosmic ray particles due to the constant ionization loss in the atmosphere, becomes positive if the differential intensity of the particles increases with energy. Therefore, any maximum in the differential spectra shifts toward lower energy with an increase of atmospheric depth because of this term. Since the peak of the energy spectrum for cosmic rays exists around 10 Gev, where the corresponding path length of pions or of kaons is only of the order of a meter, this shift of the peak in the energy spectrum with atmospheric depth

is not conceivable for pions or kaons but only for muons whose path length is of the order of several tens of kilometers.

I-3. THE THIRD APPROXIMATION—WITH RADIATION LOSS

The rate of energy loss of charged mesons of cosmic rays in the air, including the radiation loss is

$$(-dE/dx)i=a_i+b_iE_i \qquad (I.5)$$

where $b_i=r_i+p_i$ and a_i, r_i, and p_i correspond, the rate of energy loss due to ionization, due to radiation (Bremsstrahlung), and due to direct pair creation in the air, respectively.

According to Barrett et al., (1952), the value of r and p in the earth are 1.5×10^{-6} g/cm^2 and 1.3×10^{-6} g/cm^2, respectively. These values correspond to muons penetrating underground. For the following reasons, it can be assumed that r_i 1.0×10^{-6} g/cm^2 and p_i 1.5×10^{-6} for $i=\pi$ (pions) in the atmosphere:

(i) Z/A is roughly the same order both for the air (N_2 and O_2) and for the earth (SiO_2 and H_2O), whereas the radiation length which is inversely proportional to Z^2/A is slightly longer for the air than for the earth. Since the present calculations are based on the atmospheric depth, which is proportional to Z/A, r and p should be reduced by a factor which is approximately the ratio of radiation length of the earth to that of the air.

(ii) Energy losses expressed by Eq. (I.5) are only due to Coulomb interaction; therefore parameters used for muons can be used for pions since the rest masses of these particles are of the same order. For kaons b_i should be reduced by a factor of 10 which is approximately the square of the ratio of rest mass of pions to that of kaons. The constant ionization factor a_i is assumed to be 2.5×10^{-3} Gev/g cm^2 for all cosmic ray particles. Since the ranges of cosmic ray mesons are very short, compared with muons, the error due to these parameters is practically negligible. Since E_i and E_i'' are the energy of mesons at the atmospheric depths x and x', respectively from the range-energy relation

$$\int_{E_i''}^{E_i} dE_i' \Big/ \left(-\frac{dE_i'}{dx}\right) = \int_{x'}^{x} \sec\theta^*(x'')\,dx'' \simeq (x-x')\sec\theta^*(x).$$

By using Eq. (I.5)

$$E_i''=[(a/b)+E_i]\exp[b(x-x')\sec\theta^*(x)]-(a/b) \qquad (I.6)$$

Since b is very small, Eq. (I.6) may be written

$$E_i''=E_i+(a+bE_i)\cdot(x-x')\sec\theta^*(x) \qquad (I.7)$$

Thus Eq. (2.1) is

$$n_i(E_i,x,\theta^*)=\frac{\exp[-x/L_i\cos\theta^*(x)]}{L_c\cos\theta^*(x)}\int_0^x F(E_i'')\exp[-x'/L'\cos\theta^*(x)] \cdot\left(\frac{x'}{x}\cdot\frac{E_i}{E_i''}\right)^{[B_i\sec\theta^*(x)]/[E_i+(a_i+b_iE_i)x\sec\theta^*(x)]}dx' \qquad (I.8)$$

where $1/L'=1/L_n-1/L_i$.

The diffusion equation, whose solution is given by Eq. (I.8) is no longer simple because of the energy-dependent energy loss term. It should be noted, however, that below about 50 Gev, the first approximation given by Eq. (I.1) is sufficiently accurate in any direction, because at such low energies the range of mesons before the decay into muons is so short that neglecting energy loss before decay does not cause any significant error. Then the computation of Eq. (2.1) can be done analytically as follows: (M.1).†

†These expressions and corresponding curves of Fig. 15 in M.1 are not correct, because the term $\cos\theta$ in the denominator corresponding to (1.9) is misprinted as $\sec\theta$.

(i) For large values of x with

$$B_i/E_i \cos\theta = q \quad \text{and} \quad x/\cos\theta L' = t,$$

$$n_i(E_i, x, \theta) = [F(E_i)/(L_c \cos\theta)] \times \exp[-(L'/L_i)t] L' t^{-q} \Gamma_t(q+1) \qquad \text{(I.9)}$$

where Γ_t is an incomplete gamma function, i.e.

$$\Gamma_t(q+1) = \int_0^t n^q e^{-n}\, dn$$

(ii) For small values of x, by expanding the integrand in power series of x and integrating by term,

$$n_i(E_i, x, \theta) = [F(E_i)/(L_c \cos\theta)] \times \exp[-(L/L_i)t] \sum_{n=0}^{\infty} (-1)^n \frac{t^{n+1}}{n!(q+n+1)} \qquad \text{(I.10)}$$

Appendix II

ACCURACY OF EQ. (3.1.6)

The differential spectrum of cosmic ray mesons at the atmospheric depth x for the zenith angle direction θ^* with energy E_i in dE_i is empirically given by

$$n_i(E_i, x, \theta^*) = A E_i^{-m} \cdot f(x, \theta^*) \qquad \text{(II.1)}$$

Equation (3.1.6) is then

$$n_\mu(E, x, 0) = \int_{E_\mu}^{E_{\mu r^2}} \frac{B_i \sec\theta^*(x)}{E_i x} \cdot \frac{r_i^2}{r_i^2 - 1} \cdot \frac{A E_i^{-m}}{E_i} dE_i$$

$$= \frac{B_i \sec\theta^*(x)}{E_\mu x} F(m) \cdot n_i(r_i E_\mu, x) \qquad \text{(II.2)}$$

where $r_i = m_i/m_\mu$ and

$$F_i(m) = \frac{r_i^2}{r_i^2 - 1} \cdot \frac{r_i^m}{m+1} [1 - r_i^{-2(m+1)}] \qquad \text{(II.3)}$$

The values of $F_i(m)$ are shown in the following table for pion produced muons ($i = \pi$), and Kaon produced muons ($i = k$).

m	0	1	2	3	4
$F_\pi(m)$	1.00	1.04	1.09	1.19	1.33
$F_k(m)$	1.00	2.44	7.59	26.5	98.8

Since Eq. (3.1.6) is derived by disregarding a factor $F(m)$ the differential intensity calculated by this formula gives nearly 10% under-estimation for pion produced muons with $m \simeq 2$, and the error (under-estimation) is too large to calculate Kaon produced muons.

ACKNOWLEDGMENTS

I am grateful to Mr. E. Monasterski who helped with most parts of numerical integrations by machine (IBM 7090).

REFERENCES

Ascoli, G., "Relation between Pi- and Mu-Meson Production Spectra in the Atmosphere," *Phys. Rev.*, **79,** 812, 1950.

Ashton, F., "The Range-Energy Relation for High Energy Mu-Mesons," *Proc. Phys. Soc. London,* **77,** 587–592, 1961.

———, and Wolfendale, A. W., "The Momentum Spectrum of Cosmic Ray at an Angle of 80° to the Zenith at Sea Level," *Proc. Phys. Soc. London,* **81,** 593–603, 1963.

———, Brooke, G., Gardener, M., Hayman, P. J., Jones, D. G., Kisdnasamy, S., Loyd, J. L., Taylor, F. E., West, R. H., and Wolfendale, A. W., *Nature,* **185,** 364–365, 1960.

Barrett, P. H., Bollinger, L. M., Cocconi, G., Eisenberg, Y., and Greisen, K., "Interpretations of Cosmic Ray Measurements Far Underground," *Ref. Mod. Phys.*, **24,** 133–178, 1952.

Duthie, J. G., Fisher, C. M., Fowler, P. H., Kaddoura, A., Perkins, D. H., Pinkau, K., and Wolter, W., "The Study of High Energy Gamma Rays Produced by Cosmic Radiation at 40,000 Feet," Part II, "The Energy Spectrum of Cascades and Its Interpretation," *Phil. Mag.*, **6,** 113–131, 1961.

———, Fowler, H., Kaddoura, A., Perkins, D. H., and Pinkau, K., "The Flux of Rays at High Altitudes and Comparison with the Muons at Sea Level," *Nuovo Cimento,* **24,** 122–138, 1962.

Hayakawa, S., and Tomonaga, S., "Cosmic Ray Underground," *Prog. Theor. Phys.*, **4,** 287–296, 1949.

Hayman, P. J., Palmer, N. S., and Wolfendale, A. W., "Range Straggling of Muons in the Energy Range 200–100,000 Gev," *Proc. Phys. Soc. Lond.*, **80,** 800–802, 1962.

Maeda, K., "Directional Dependence of Atmospheric Temperature Effects on Cosmic Ray Muons at Sea Level," *J. Atmospheric Terrest. Phys.*, **19,** 184–245, 1960.

———, and Wada, M., "Atmospheric Temperature Effect upon the Cosmic Ray Intensity at Sea Level," *J. Scient. Res. Inst. Tokyo,* **48,** 71–79, 1954.

Moroney, J. R., and Parry, J. K., "Momentum Distribution of Mu-Mesons at Zenith Angles in the East-West Plane," *Australian J. Phys.*, **7,** 423–438, 1954.

Ozaki, S., "On the Energy Loss of the Mu-Mesons in the Ground," *J. Phys. Soc. Japan,* **17,** Suppl. A III (Intern'l Conf. Cosmic Rays and Earth Storm Part III), 330–332, 1962.

PAK, W., OZAKI, S., ROE, B. P., and GREISEN, K., "Momentum Spectrum of Muons," *Phys. Rev.*, **121,** 905–907, 1961.

RODGERS, A. L., "On the Momentum Spectrum of Cosmic Ray Muons at Sea Level," *Proc. Phys. Soc. London*, **78,** 918–925, 1961.

SANDS, M., "Low Energy Mesons in the Atmosphere," *Phys. Rev.*, **77,** 180–193, 1950.

SHELDON, W. R., and DULLER, N. M., "Intensities of Muons Underground at Large Zenith Angles," *Nuovo Cimento*, **23,** 63–76, 1962.

SMITH, J. A., and DULLER, N. M., "Effects of Pi Meson Decay-Absorption Phenomena on the High-Energy Mu Meson Zenithal Variation near Sea Level," *J. Geophys. Res.*, **64,** 2297–2305.

WADA, M., "Atmospheric Effects on the Intensity of Cosmic Ray Mesons I, The Barometers Effect," *Sci. Papers Inst. of Phys. Chem. Res., Tokyo*, **54,** 335–352, 1960.

ZATSEPIN, G. T., and KUZ'MIN, V. A., "Angular Distributions of High-Energy Muons in the Atmosphere and their Production Mechanism," *Soviet Phys. JETP*, English Translation, **12,** 1171–1177, 1961.

CERENKOV-SCINTILLATION COUNTER MEASUREMENTS OF THE LIGHT, MEDIUM, AND HEAVY NUCLEI IN THE PRIMARY COSMIC RADIATION FROM SUNSPOT MINIMUM TO SUNSPOT MAXIMUM*

F. B. McDONALD AND W. R. WEBBER†

Goddard Space Flight Center

The intensity and rigidity spectra of Be, B, C, N, and O have been measured on a series of three Skyhook balloon flights using a Cerenkov-scintillation detector combination. The flights occurred at times close to solar minimum and maximum. The energy and charge resolution is comparable with that obtained by the latest emulsion techniques. It is observed that the light and medium nuclei display the same relative rigidity spectra throughout the solar cycle as the protons and α particles. This conclusion is valid only in the region studied ($E>400$ Mev/nucleon), and we may well expect striking variations for heavy primaries of low energy. The measured ratio of light to medium nuclei at the top of the atmosphere was 0.36 ± 0.06. This was not observed to vary appreciably over the solar cycle.

INTRODUCTION

Measurements on primary cosmic-ray nuclei heavier than α particles have provided much information helpful in understanding the problems relating to the acceleration and the propagation of primary cosmic rays to the earth. Two characteristic features of these heavier nuclei are utilized in interpreting these measurements.

1. Because of their fragile structure, which can be easily destroyed in nuclear interactions at cosmic-ray energies, heavier nuclei cannot be created from lighter elements once they are ejected from the source into interstellar space. Thus any conceivable process the primary cosmic radiation might undergo in interstellar space takes place in the direction from heavier to lighter elements. Experimentally this feature is reflected in the study of the relative abundance of each of the heavier nuclei and, in particular, the relative abundance of the so-called light nuclei (Li, Be, B), which are commonly believed to be absent in the cosmic-ray source regions. A knowledge of these relative abundances, together with the fragmentation parameters of these nuclei in the interstellar medium, can lead to definite limitations on acceleration and propagation processes of the primary cosmic radiation.

2. The rate of energy loss in the passage through interstellar material is greater for the heavier nuclei, being proportional to the Z^2/β^2 of the particle. The effect of this ionization loss will be most pronounced at the low-energy end of the spectrum. Experimentally this feature should be revealed by a comparison of the differential energy spectra of the various charges, particularly at energies where β is appreciably less than 1. Thus a knowledge of the shape of the differential spectra for the various charges can place further limitations on cosmic-ray acceleration processes and

*The balloon flights from which the data reported here were obtained were carried out while the authors were at the State University of Iowa, Department of Physics, Iowa City, Iowa, and were sponsored by the Office of Naval Research.

†Now at the School of Physics, University of Minnesota, Minneapolis, Minnesota.

on the travel of cosmic rays in the interstellar medium.

Recent progress in balloon techniques and the practicality of studies in satellites have opened the possibility for further advances in the study of these heavier nuclei. Particularly, it is now possible to study the energy spectra of the individual nuclei or groups of nuclei in much the same manner as proton and α-particle spectra have been studied for a number of years. The additional information to be derived from such studies, as is noted above, opens up a new dimension for attacking the problems outstanding in this field.

A number of preliminary investigations of the low-energy portion of the spectra of the heavier nuclei ($Z \geqq 6$) have recently been made by means of emulsion techniques. They have led to a somewhat confused picture however. For example, the results reported by Fichtel[1] and Evans[2] tend to suggest that the differential spectra of the light L, medium M ($6 \leqq Z \leqq 9$), and heavy $H(Z \geqq 10)$ nuclei are similar to those for protons and α particles (when the intensities are expressed in particles/m^2-ster-sec-Bv). Aizu et al., from results of much greater statistical significance,[3,4] find an anomalous spectrum for the L nuclei. Finally, Tamai[5] reports systematically different spectra for all charge groups, the peak in the differential spectrum occurring at higher energies for the particles of higher Z.

It should be pointed out that all these measurements were made at times when the solar modulation of cosmic rays was appreciable and variable, that is, at times near *sunspot maximum*. The effect of this modulation on the heavier nuclei has not been directly measured on separate occasions over the sunspot cycle but has been inferred from comparison of these spectra with those of protons and α particles reported at similar times. Since it has been established that the solar modulation for protons and α particles is similar when expressed in terms of changes in the rigidity spectra of these particles,[6,7,8] we should expect this rigidity dependence to hold for the heavier nuclei as well because of their identical Z/A.

In actual fact, because of the solar modulation effects, which are certainly large and not yet known with sufficient accuracy, it is necessary to measure the spectra of the heavier nuclei at times near *sunspot minimum* when solar modulation effects are probably small, or at least do not fluctuate rapidly—that is, if we hope to obtain useful information on the acceleration and propagation of the primary cosmic radiation outside of the solar system by these measurements.

In the light of the above situation, and to investigate the ability of counter techniques to examine the charge and energy spectra characteristics of the heavy nuclei, it seems worthwhile to report here the results from a series of three flights using Cerenkov-scintillation counters to measure the energy spectra of charges with $3 \leqq Z \leqq 10$. The data were obtained in conjunction with a series of more than twenty flights to measure the proton and α-particle spectra during the period 1954 to 1959.[6-10] The three flights cover the period from sunspot minimum to sunspot maximum. The flight near sunspot minimum provides the only spectral measurements of the heavier nuclei yet reported and as such will be used in deducing certain new limitations on the propagation and acceleration of the galactic radiation. The measurements of the cumulative effects of the solar modulation on the heavier nuclei will be used to extend the conclusions reached earlier for the effects of this modulation on protons and α particles.

DESCRIPTION OF DETECTOR USED

The detector used in the measurements reported here is identical to that used to measure protons and α particles; the detector and recording system have been described in detail previously.[9] Briefly, the detector is a three-element telescope consisting of an NaI crystal scintillation counter, a Lucite Cerenkov counter, and a tray of Geiger counters (except for third flight; see below). The scintillation crystal and the Geiger counter tray are the defining elements of the telescope. For each particle that traverses these elements, the outputs from the Cerenkov counter and the scintillation counter are recorded. The relation between the measured outputs from the Cerenkov counter ($I_c \approx Z^2)(1-1/\beta^2 n^2)$ and the scintillation counter ($I_s \approx Z^2/\beta^2$) should uniquely determine the charge and velocity of the particle in question. In actual

practice it is possible to measure the spectra of the heavier nuclei in the range of $\approx$250 to 1500 Mev/nucleon. The low-energy limit, which is Z-dependent, is determined by the ionization cutoff introduced by the material in the detector and the atmosphere above, and also by general background pulses in this region of zero Cerenkov output. A system of guard counters is used to detect and minimize such background counts, usually due to interactions and knock-on electrons occurring in the detector and particularly in the detecting elements themselves.

The instrumental calibrations are a very important aspect of this experiment. It is necessary first to identify the pulse height distributions of the relativistic particles of each charge and then to relate these positions to the output pulses from the lower energy particles in order to ascertain their energy. The response and linearity of the electronic system were checked carefully before and after each flight by introducing simulated photomultiplier pulses whose amplitude could be controlled accurately as multiples of Z^2, where $Z=1$ amplitude was established to within 5 percent by tests using μ mesons at sea level.

In practice the ground calibration is confirmed by noting the relative positions of the peaks of the ionization and Cerenkov distributions of relativistic α particles and carbon nuclei at balloon altitudes. The centers of these distributions corresponding to $(I_s)_{\min}$ and $(I_c)_{\min}$ can be determined to an accuracy of ± 10 percent. These positions agree to within the experimental error with those expected from a linear system on the basis of the sea-level calibration. The positions $(I_s)_{\min}$ and $(I_c)_{\min}$ for the other charge components can be determined with less accuracy but do not indicate any nonlinearities in the system. From this we conclude that saturation effects are small in NaI crystals for relativistic particles with $Z \leqq 10$.

The energy spectra of the heavier nuclei are determined by using the theoretical variations of I_s and I_c with velocity. It is estimated that energies in the range 250–1500 Mev/nucleon can be measured to an accuracy of $\approx$10 percent of less for particles with $4 \leqq Z \leqq 10$. The observed counts appear to fall along the predicted curves for these lower energy particles, and there is no evidence for a systematic distortion of the energy calibration.

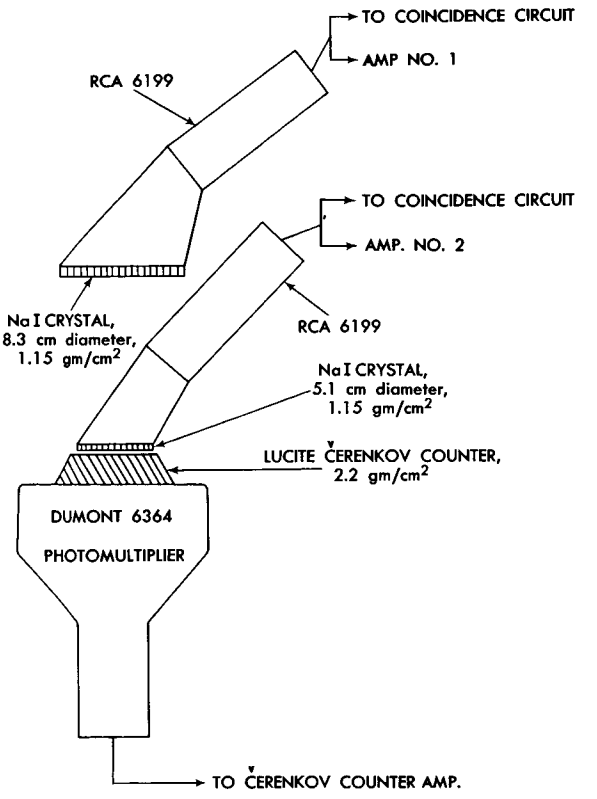


FIGURE 1.—Double-scintillation—Cerenkov telescope. For an event, the pulse height of each of the three counters is recorded.

The third flight reported here was made with a somewhat different detector: A second NaI scintillation crystal was used in place of the Geiger counter tray. The two scintillation crystals were the defining elements of the telescope, and for each particle three outputs (two scintillation and one Cerenkov) were recorded—giving, in effect, a double-scintillation—Cerenkov detector. A schematic drawing of this detector is shown in figure 1. Its characteristic features are generally similar to the earlier one, but the replacing of the Geiger counter tray by the second scintillation crystal led to certain notable improvements: (1) The second scintillator provided a self-consistent check on the charge and energy of the particles as determined from the relative outputs of the Cerenkov counter and the other scintillation crystal; comparison of the outputs of the two scintillation crystals gave

TABLE 1.—*Details of Balloon Flights*

Flight Data	Flight Number		
	6	7	A
Date	March 20, 1956	August 17, 1956	August 1, 1958
Altitude (average)	6.8 gm/cm²	5.3 gm/cm²	5.0 gm/cm²
Location	Iowa City, Iowa	Minneapolis, Minn.	Moberly, Mo.
Geomagnetic threshold during flight*	1.7−1.2 Bv	1.2±0.1 Bv	2.5±0.1 Bv
Hours of data	6	4	6
Geometric factor	7.2 ster-cm²	7.2 ster-cm²	6.4 ster-cm²
Amount of absorber in detector	11.4±2 gm/cm² (unit 1)	11.4±2 gm/cm² (unit 1)	8±2 gm/cm² (unit 2)
Mt. Washington neutron intensity at time of flight (scaled counts per 2-hr interval).†	2205	2415	1975

*After Quenby and Webber (reference 12).
†Courtesy of Prof. J. A. Lockwood, University of New Hampshire.

an additional parameter to check the change and energy, namely, a measurement of the rate of energy loss of the particles at two different points on the trajectory. (2) The second scintillator also reduced the problem of background counts to negligible proportions. Since most of these "confusion" counts were due to interactions or knock-on electrons produced in the material of the detector itself, a greater degree of selection for these local events could be made by comparing the two scintillation outputs that reflected ionization loss conditions upon entrance and exit from the detecting system. Only events consistent with single particles traversing the system were used. The extra scintillator acted as a "guard" system so efficiently that it was possible to dispense with the "guard" counters in the later system.

EXPERIMENTAL RESULTS

The data on which this report is based, as noted before, come from three balloon flights of the Cerenkov-scintillation or Cerenkov-double-scintillation detector (see Table 1). As a measure of the cumulative effects of solar activity at the time of the flights the average Mt. Washington neutron monitor hourly scaled counting rates are given for the times the detectors were at peak altitude. The peak hourly scaled rate at Mt. Washington during the period of actual sunspot minimum in 1954 was ≈2500; thus conditions on August 17, 1956, indicate a solar modulation effect of ≈4 percent as recorded by a high-latitude neutron monitor. On March 20, 1956, the intensity was recovering from a Forbush decrease, and the total solar modulation was ≈12 percent. Finally on August 1,

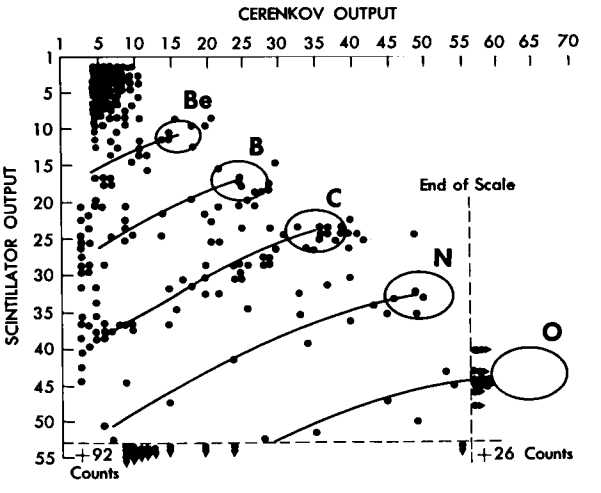


FIGURE 2.—Two-dimensional plot of Cerenkov light output vs. ionization loss for counts recorded on flight 6, March 20, 1955. The complete absence of counts in the region of high Cerenkov pulse height and the small scintillator output indicate the lack of background in the fast particle region. The lithium region is obscured by background, and there are probably background counts in the region below 400 Mev/nucleon.

1958, the cumulative solar modulation was ≈22 percent, which represents almost the full magnitude of the solar modulation effects as recorded by such a monitor.[11]

Figures 2, 3, and 4 show the two-dimensional plots of all the counts recorded on the three flights that were due to single particles, with $Z>3$, that traversed the detectors. The theoretically determined curves for the output of the detectors as a function of energy and charge are shown as solid lines. The normalization procedure is that discussed in the previous section. In figure 4, which represents the results of the flight using a Cerenkov-double-scintillation counter, both scintillation outputs are shown, a single particle being assumed to traverse the telescope without interaction if the outputs of the two scintillators are each within ±20 percent of their mean value.

It should be pointed out that the intensities of protons and α particles were also measured on these flights. This was accomplished by switching the gain of the electronics so that for 15 minutes of every hour the gain was increased by a factor of 8 so as to record the smaller proton and α-particle outputs. We do not analyze the proton and α-particle results here in any detail but only use them for comparison with the heavier particle results on the same flights.

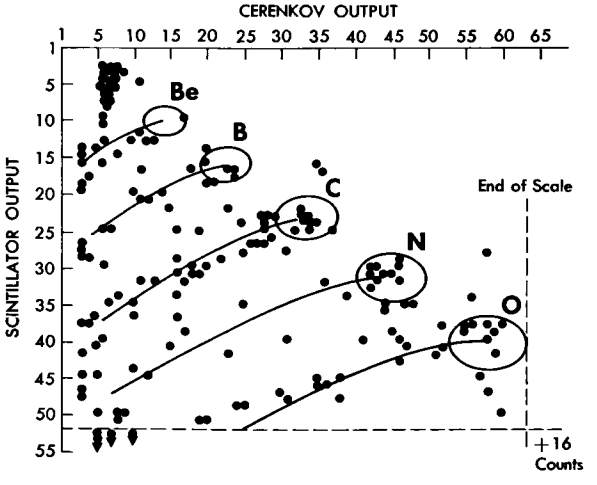


FIGURE 3.—Two-dimensional plot of Cerenkov light output vs. ionization loss for counts recorded on flight 7, August 17, 1956. As in figure 2, the curves indicate the theoretically expected distribution. The separation between boron and carbon is particularly good for both flights 6 and 7.

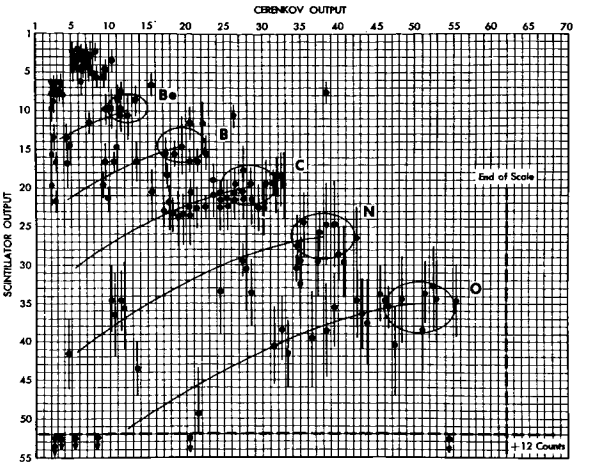


FIGURE 4.—Plot of Cerenkov pulse height vs. ionization loss in the three-element telescope of figure 1. The end points of the individual lines represent the two measurements of ionization loss. A selection criterion that demands that the two measurements be within ±20% of each other has been applied. An appropriate correction for Landau fluctuations is then made. There is no evidence for background in the region beyond lithium, and the charge resolution is uniformly excellent.

The charge and energy resolution of this type of detecting system for the heavier particles is evident from the figures without need for further elaboration. However, to convert the data contained in the figures to actual intensity and differential spectra values for the various charge components, a number of corrections, discussed in detail in the following section, are necessary.

DETERMINATION OF ABSOLUTE INTENSITIES OF VARIOUS NUCLEI

Since we have selected only counts that are clearly due to single particles traversing the detecting system, it is necessary to make a number of corrections to the raw data reported in the previous section. These corrections are not large and, furthermore, are reasonably well known. They are very similar to the corrections that were used to determine the absolute flux of protons and α particles in a concurrent series of balloon flights and were discussed in detail by McDonald[9] and McDonald and Webber.[6]

For the heavier nuclei the data in figures 2 through 4 must be corrected (1) for the nuclear interactions of the heavier particles in the de-

TABLE 2.—*Corrections to Raw Balloon Flight Data*

Corrections (percent)	Flight		
	6	7	A
Correction for nuclear interactions in telescope*:			
λ_L' (glass) = 45 gm/cm²	−28±5	−28±5	−19±5
λ_M' (glass) = 40 gm/cm²	−33±6	−33±5	−24±6
λ_H' (glass) = 30 gm/cm²	−43±8	−43±8	−30±6
Knock-on electron correction for $E>0.5$ Bev/nucleon:			
L	−2±1	−2±1	
M	−5±3	−5±3	None
H	−12±8	−12±8	
Correction to the top of the atmosphere* (diffusion extrapolation):			
L	−0±5	−0±5	−0±5
M	−24±5	−19±4	−18±4
H	−30±6	−25±5	−24±5

*Interaction mean free paths in glass and air, and fragmentation parameters, after Waddington (reference 13).

tecting system itself, resulting in the absorption of these particles; and (2) for the effects of knock-on electrons produced by these particles as they pass through the detector. Finally, to obtain the absolute intensities at the top of the atmosphere, it is necessary to make a diffusion extrapolation of the corrected intensities at the flight altitudes to the top of the atmosphere.

TABLE 3.—*Absolute Intensities of the Various Components at the Top of the Atmosphere*

Flight number and date	Energy interval at top of atmosphere (Bev/nucleon)	α Particles m²-ster-sec	Light nuclei* m²-ster-sec	Medium nuclei m²-ster-sec	L/M	Heavy nuclei† m²-ster-sec
6: Mar. 20, 1956	>0.43	176±6	4.3±0.7	10.2±0.9	0.42±0.06	3.4±0.6
	0.43–0.57	19.3±1.3	1.1±0.4	1.7±0.4	0.65±0.19	
	0.57–1.06	45.5±3	1.2±0.4	2.6±0.4	0.46±0.15	
	1.06–1.55	24.3±2	0.45±0.2	1.8±0.4	0.25±0.09	
	>1.55	86±3	2.1±0.5	4.4±0.6	0.48±0.09	
7: Aug. 17, 1956	>0.41	225±10	4.5±0.9	15.3±1.3	0.29±0.08	3.8±0.8
	0.41–0.55	23±3	0.7±0.3	2.3±0.5	0.31±0.12	
	0.55–1.04	67±5	1.8±0.6	4.4±0.7	0.41±0.12	
	1.04–1.53	36±4	0.8±0.4	2.4±0.5	0.33±0.14	
	>1.53	88±5	1.3±0.6	6.4±0.9	0.2±0.06	
A: Aug. 1, 1960	>0.55	105±6	2.5±0.5	7.2±0.6	0.35±0.07	1.8±0.4
	0.55–1.04	20±2	0.8±0.3	1.8±0.4	0.45±0.12	
	1.04–1.53	14±2	0.2±0.1	1.1±0.4	0.18±0.12	
	>1.53	12±1	1.7±0.4	4.1±0.5	0.41±0.10	

*Li $1/4L$.
†$Z \geq 10$.

1. *Correction for Nuclear Interactions in the Detector*

Since the raw data include only clearly identifiable single particles passing through the telescope, they exclude those particles that interact as they pass through the telescope. On the reasonable assumption that at least two products occur for each interaction, these events should be recorded by the guard system (or, with the second detector, the pulse height after the interaction is significantly different than that before). Subsidiary experiments[6] suggest the guard system has in fact an efficiency of ≈ 1 for such events. The correction for the number of counts lost because of nuclear interactions can then be made, the interaction mean free paths of the various components in the telescope being known. This correction is given in Table 2.

2. *Correction for Knock-on Electrons Produced in the Detector*

It is possible for particles, as they pass through the telescope, to produce knock-on electrons having sufficient energy to set off at least one of the guard counters. Such events will be rejected, even though a single particle traverses the telescope without interaction. This correction depends on the charge and energy of the particle in question. Detailed calculations of this effect have been made to interpret previous experiments on protons and α particles[9,14] This correction, as applied to the heavier nuclei in this experiment, is shown in Table 2. Note that no correction for this effect is necessary in the case of the Cerenkov-double-scintillation counter array.

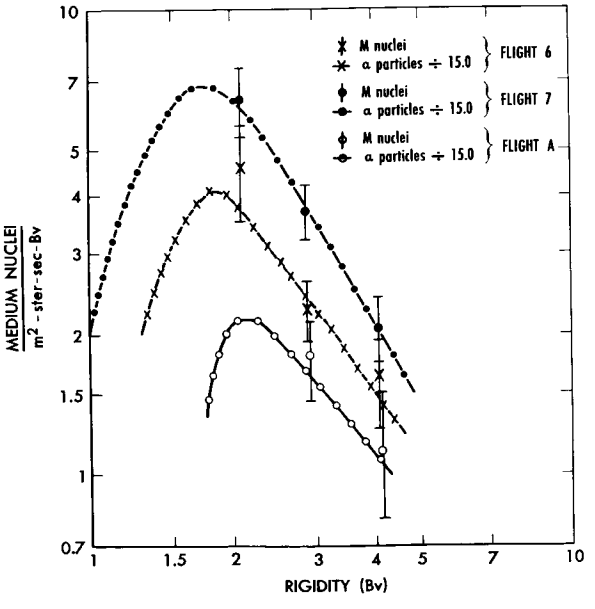


FIGURE 5.—Measured differential rigidity spectrum of M nuclei. The appropriate α-particle data divided by 15 are also shown for comparison.

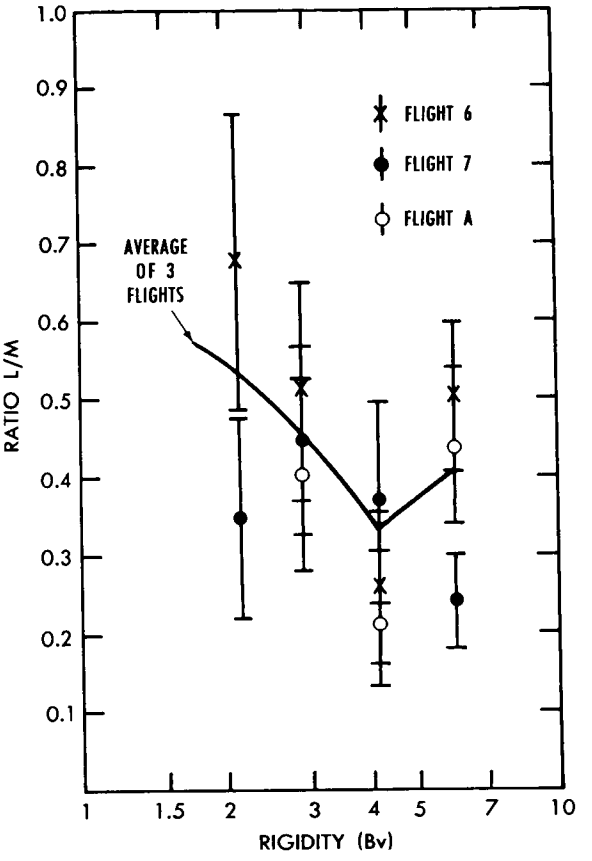


FIGURE 6.—The measured L/M radio as a function of rigidity. In the region of these studies, $1.5 \leqq R \leqq 6$, the results are consistent with a constant L/M over the entire rigidity interval, although there is some evidence it may be increasing at smaller values of R.

The absolute flux values obtained after the above corrections are extrapolated to the top of the atmosphere in the conventional manner with the diffusion equations as introduced by Kaplon, Noon, and Racete.[15] The interaction mean free paths and fragmentation probabilities used are those determined by Waddington[13] in his survey of available data on the subject, and it has been assumed that they do not vary appreciably over

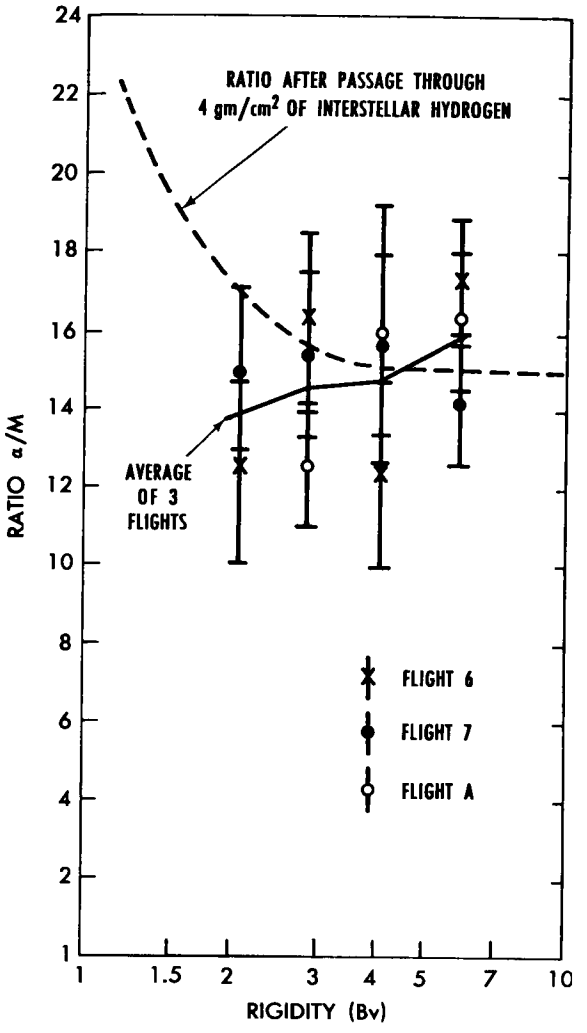


FIGURE 7.—The measured α/M ratio as a function of rigidity. The dotted line indicates the expected ratio after passage through 4 gm/cm² interstellar hydrogen, assuming that originally α and M nuclei have the same form of rigidity spectrum.

this range of energies. Energy loss has not been introduced directly into the diffusion equations; instead, the energy intervals appropriate to the flight altitude $\approx$6 gm/cm² for each charge group have been corrected for energy loss to give the equivalent intervals at the top of the atmosphere. This will give a small but insignificant bias due to wrong assessment of energies.

Table 3 summarizes the results of the three flights corrected to the top of the atmosphere. Figure 5 shows the differential spectra of M nuclei that were measured on each of the flights. For comparison the differential spectra of α particles measured on the same flights and divided by a constant factor of 15 are shown. Finally, in figures 6 and 7 the L/M ratio and α/M ratio are shown as a function of energy for the average of the three flights and for each flight separately.

DISCUSSION OF EXPERIMENTAL RESULTS

We shall divide our discussion of the measurements reported in the previous section into three separate aspects:

1. *The Simple L/M Ratio*

The average L/M ratio at the top of the atmosphere for the three flights is 0.36 ± 0.06 for particles with energy greater than 1.55 Bev/nucleon or 0.38 ± 0.05 for particles with energy greater than 0.41 Bev/nucleon. These values are in generally good agreement with those that have been obtained by emulsion techniques. In his survey of all previous determinations using these methods, Waddington[13] derives a mean L/M ratio of 0.29 ± 0.03 for particles with energy greater than 1.5 Bev/nucleon. The more recent L/M determinations of Van Heerden and Judek,[16] 0.35 ± 0.04, and O'Dell et al.,[17] 0.25 ± 0.05, for particles with energies greater than 1.5 Bev/nucleon—or Aizu et al.,[3,4] 0.41 ± 0.06, and Fichtel,[1] 0.40 ± 0.08, for particles with energies greater than 0.4 Bev/nucleon—give a clear indication that there is no longer any substantial uncertainty in this ratio, whether it is measured by emulsion techniques or by counter techniques.

A clear difficulty arises in using this ratio to attempt to calculate the amount of interstellar material traversed by the primary radiation however. Making the usual assumption that the L nuclei are absent in the source region and the additional, and artificial, assumption that all particles traverse the same amount of interstellar material (one-dimensional diffusion), Hayakawa et al.[18] have concluded that an L/M ratio of 0.3 implies that the most probable amount of matter traversed was 3 ± 0.5 gm/cm². Using more recent values for the fragmentation parameters of the heavier nuclei in hydrogen, Fichtel[19] arrived at a value of 5 ± 2 gm/cm² for the same L/M ratio. The

differences in these values and the uncertainties in the values themselves represent the degree of our uncertainty about the fragmentation parameters for heavy-nuclei collisions with interstellar hydrogen.

A number of other possibilities exist for models of the diffusion of the primary radiation through interstellar space—for example, steady-state three-dimensional diffusion from a point source in an infinite medium of constant density; the same except with uniform injection throughout the volume and either slow leakage or perfect reflection at the boundary.[20,21] In these models the L/M ratio is a function of energy (even without considering the effects of Fermi acceleration or ionization loss in interstellar space). We shall discuss this fact further in the following sections. If the cosmic rays are restricted to lines of force, which may be more regular in the galaxy than is sometimes supposed, then the usual diffusion treatments must be modified. The direct injection of L nuclei by spallation in stellar atmospheres where heavier cosmic rays have already been accelerated may also influence the interpretation of the L/M ratio.

2. *The Solar Modulation Effects on the Heavier Nuclei*

The work of McDonald,[9] McDonald and Webber,[6,7] Webber,[10] and Meyer[22] has established to quite a high degree of accuracy that the sunspot cycle changes or modulation of the proton and α-particle intensity are identical for particles of equivalent rigidity. Furthermore, the intensity of protons or α particles at the top of the atmosphere can be uniquely related to the counting rate of a sensitive neutron monitor near sea level at the same time. Since the heavier nuclei have effectively the same charge to mass ratio as the α particles, we should expect the rigidity-dependent modulation to be effective for them also. Because of the approximately similar ratio of Z/A for α particles through oxygen, it is impossible to distinguish between rigidity-dependent modulation and modulations involving energy/nucleons without considering the proton data. However, since the proton and α-particle data have established a rigidity-dependent modulation, the interpretation of the modulation of light and medium nuclei is carried out using rigidity. While

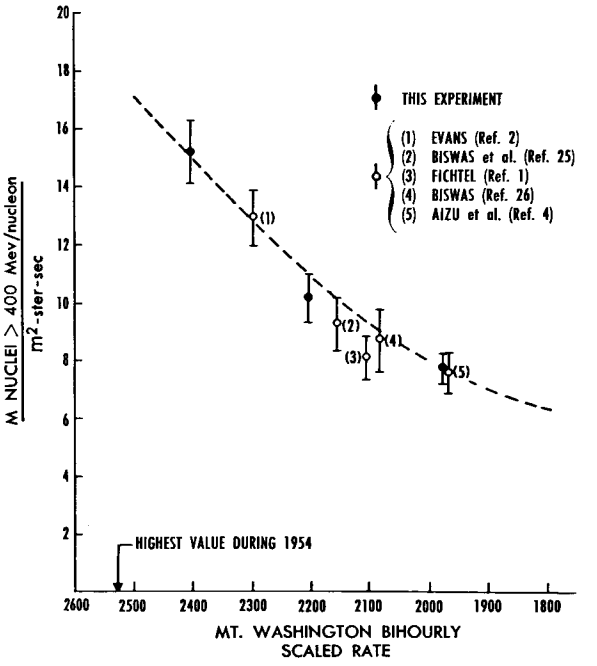


FIGURE 8.—The variation of integral medium nuclei intensities above 400 Mev/nucleon during the period 1956–59 as measured by various observers. Intensities are shown as a function of the Mount Washington neutron counting rate at the time of measurements. The dotted curve represents McDonald and Webber's (reference 7) proton and α-particle results divided by 72 and 12, respectively.

it would be more convincing to plot L and M with proton data, the α-particle flux values are used since greater overlap between the two sets of data is obtained during a given flight. Observations by Frier et al.,[23] Fichtel,[1] Van Heerden and Judek,[16] Yagoda,[24] and Biswas et al.[25] during the recent sunspot maximum period have clearly indicated that the intensity of heavier nuclei varies considerably throughout the sunspot cycle and is much lower at this time than at sunspot minimum. This variation has been directly related to the concurrent variations of protons and α particles by the above authors.

The results included here cover an extended period from a very high cosmic-ray intensity indicative of sunspot minimum to a low cosmic-ray intensity indicative of sunspot maximum. Perhaps the best way to illustrate the behavior of the heavier nuclei during this period is to plot the intensity of the most abundant of these nuclei, the

M nuclei, against the counting rate of a sensitive neutron monitor. Such a plot is shown in figure 8. Since deviations from a rigidity-dependent modulation would be expected to be most pronounced at lower energies, we have included only measurements that extend down to $\approx$500 Mev/nucleon or less. The dashed curve in figure 8 represents the intensity of α particles $>$400 Mev/nucleon as a function of the Mt. Washington neutron monitor rate as deduced from the α-particle measurements on these three flights, as well as a more extended analysis by McDonald and Webber[7] of proton and α-particle data over twenty flights during the sunspot cycle. The α-particle intensities are divided by a constant factor of 15 ± 1. There is no suggestion from the figure that anything but a constant factor suffices to normalize the proton and α-particle and M-nuclei data over the sunspot cycle despite the fact that the integral intensity of M nuclei varies by a factor of greater than 2 during this time.

Examination of the differential spectra of α particles and M nuclei derived from the three flights and shown in figure 5 also suggests that there are no systematic changes in the relative differential spectra of these charge groups during the solar cycle.

The above arguments may be extended with less accuracy to include the L and H nuclei as well. No systematic changes in the L/M or H/M ratios are noted during the course of the three flights (see Table 3 and figure 6).

The results of the heavier nuclei are thus consistent with the modulation of the α particles. When taken with the proton–α-particle modulation characteristics already pointed out, they provide an additional (although statistically much less significant) support for the earlier conclusion that the solar modulation is identical for particles of the same rigidity. In particular the experimental results appear to be accurate enough to rule out simple types of electric field as the cause of the modulation of the primary radiation (e.g., reference 27).

To detect possible deviations from the simple rigidity-dependent picture to which the experimental evidence seems to be leading will require a definite advance in technique and a substantial increase in counting rates. In addition particular attention should be paid to the changes in intensity of the lowest rigidity particles (those below about 1.5 Bv), for it is this rigidity region in which the greatest deviations between the charge components might be expected.

3. *Spectral and Charge Characteristics of the Unmodulated Galactic Heavier Primary Radiation*

To assess these characteristics of the primary spectrum properly, it is necessary to be assured that we are indeed measuring the true galactic spectrum, unmodulated by any solar influences. Certainly at times other than those near sunspot minimum the solar modulation effects will not be small, particularly at low energies. It is clear from the discussion in the previous section that the features of the solar modulation have not been accurately enough delineated, so that considerable uncertainty is introduced in interpreting measurements made near sunspot maximum that have been corrected for these effects. At sunspot minimum it is likely that the solar modulation is small even for the lowest rigidity particles.[10] For this reason the results of flight 7 of this series are of particular interest.

The differential spectrum of M nuclei obtained on this flight and shown in figure 5 is of value even though the measurements do not extend to rigidities below the peak in the spectrum. The constant factor of $\approx$15, which appears to relate this spectrum to the α particle (and proton) spectrum at a given rigidity that were measured at the same time, suggests that the galactic influences on the shape of the spectra are similar to those for α particles and protons.

To investigate this feature in another way, we have plotted the α/M ratio for the three flights in figure 7. If we suppose that the modulation effects are indeed rigidity-dependent, then the results of the other two flights may be utilized with a proper degree of caution in this analysis.

The results of all of the flights are consistent with a constant α/M ratio. In fact this ratio decreases slightly with decreasing rigidity, although this decrease cannot be regarded as statistically significant. As a particular example of a set of circumstances in which the α/M should be an increasing function of rigidity towards lower lower rigidities, let us consider cosmic rays emitted

from a source region with each charge having the same differential rigidity spectrum. If we assume no further acceleration in interstellar space, the subsequent spectral shape will be determined principally by ionization loss in the interstellar material, and the heavier nuclei at a given rigidity will lose energy at a greater rate. As a consequence, differences will appear in the spectra of the low-energy heavier nuclei; and, in particular, the peaks of the differential spectra will occur at higher rigidities for particles of increasing charge.[10,28]

With these assumptions it can be seen from figure 7 that, despite the crudity of the experimental results reported here, an upper limit of ≈ 4 gm/cm^2 of interstellar material is obtained. Measurements of this type do not, of course, uniquely determine the amount of material traversed. The possibility exists that the source regions could preferentially emit the low-rigidity heavier nuclei (e.g., reference 29) so as to precisely offset the increased energy loss through ionization in the interstellar medium. Another possibility is that only processes in interstellar space are of major importance in shaping the spectra, and the competition between an acceleration process (e.g., the Fermi type) and ionization loss determines the shape of the spectra at low rigidities for the various charges. Then a critical rigidity would exist at which the rates of energy gain and energy loss by these two processes are equal. For nonrelativistic particles this rigidity can be shown to be equal to $K(A^2/Z)^{1.3}$, where K is a constant independent of the particle in question. In this picture we should also observe a systematic increase in the rigidity of the peak of the differential spectra for particles of increasing charge and, in particular, an α/M ratio that increases even more rapidly at low rigidities.

A wide variety of possibilities are therefore available for the interpretation of the results so far discussed, and it seems worthwhile to introduce the other experimental evidence relating to this matter. In figure 6 we have shown the L/M ratio as a function of rigidity as determined on the three flights. Neither the average of the three flights nor flight 7, in particular, can be interpreted as implying anything but a constant L/M ratio of ≈ 0.4 between 2 and 6 Bv. The suggestion of an increasing L/M ratio with decreasing rigidity is not statistically significant. Thus we are led again to conclude that the galactic spectra of the various charge components are closely similar when plotted as a function of rigidity.

Additional, although more indirect, evidence comes from an examination of the H/M ratio or, what is equivalent in our arguments, the α/H ratio. Any changes in the relative intensities of the different components of the galactic radiation should be most clearly shown by comparing these two groups with greatly differing charge. Although in this experiment we do not actually measure the rigidity spectrum of the H nuclei, we do obtain the integral intensity of H nuclei >400 Mev/nucleon. From this the α/H ratio turns out to be 62 ± 12 for flight 7 alone, or 57 ± 5 for the average of all of the flights. At a cutoff energy of 1.5 Bev/nucleon, the α/H ratio as deduced from the summary of Waddington, which includes the best experimental determinations up to 1959, is 42 ± 4. From this comparison we see the indication of an increasing α/H ratio with decreasing rigidity; however, this will be better established only by actually measuring the spectra of the heavier nuclei with a single instrument.

From our results on the comparative differential rigidity spectra of the various charge components we can say that, if processes such as Fermi acceleration and ionization loss in interstellar space are unimportant, the diffusion models that predict that the amount of interstellar material traversed is constant with energy are most consistent with our results. A model in which all or most nuclei of low charge are fragments of heavy nuclei that are assumed to be the only ones emitted from the source region[29,30] would not be consistent with our results. In such a picture the H nuclei would be "older" than the others and would have traversed ≈ 2 times as much interstellar material. Since the amount of material traversed must be ≈ 10 gm/cm^2 at least to produce the required fragmentation into lighter particles, this should produce a difference in spectral shape greater than that implied by the measurements.

The general picture we have tried to present above, namely, the strong similarity of the differential spectra of the various charge groups, even to the point of making the effects energy

TABLE 4.—*Comparison of Charge Distributions Obtained at the Top of the Atmosphere*

Charge	Counts*	Charge distribution (percent)						
		(†)	Waddington (ref. 13)	Aizu et al. (refs. 3 and 4)	Tamai (ref. 5)	Fichtel (ref. 1)	Biswas et al. (ref. 25)	O'Dell et al. (ref. 17)
Li	—	(6–8)	3.9	8.8	10.0	7.4	(2.1)	5.3
Be	46	6.7	1.7	6.0	14.0	5.7	(3.0)	2.3
B	70	10.1	11.6	10.9	15.7	9.0	17.8	7.4
C	197	28.6	26.0	29.2	18.8	27.1	20.9	30.1
N	92	13.3	12.4	14.8	7.8	15.3	16.6	9.7
O	123	17.9	17.9	14.4	7.3	14.4	8.9	19.4
$Z>10$	114	16.6	23.9	18.0	20.5	21.7	31.4	23.4

*Total of three flights.

†Percentage of all particles with $Z>3$ in this experiment.

loss in interstellar hydrogen on the borderline of detection, is consistent with the data of Fichtel.[1] It agrees only in part with the work of Aizu et al.[3,4,31] however. Apart from a general similarity of the spectra of the various charge groups, these authors find an L/M ratio that appears to increase with decreasing energy and that accordingly is said to reflect an increased amount of matter traversed by the low-energy particles. In addition, they find an H/M ratio that appears to decrease with decreasing energy.

On the other hand, Tamai[5] finds a systematic difference in the differential spectra of the various charge groups. The position of the maximum in the different spectra tends toward higher energy with increasing charge—an effect suggesting that ionization loss in interstellar space is an important feature. Tamai also observes an L/M ratio of ≈ 1 for particles with energy <700 Mev/nucleon, which would imply however that the elongation of the path length or the energy dependence of the fragmentation parameters dominates for the L nuclei.

From the above results it is clear that a consistent picture does not exist between various experimenters with regard to the actual shape of the low-energy portion of the galactic spectrum for the various charge groups. Two of the most serious causes for the discrepancies that now exist would seem to be: (1) *The influence of the solar modulation effects*: The three emulsion experiments reported above were all made at a time near sunspot maximum, and sometimes when the effects of Forbush decreases were important. Our present knowledge of the solar modulation effects on the heavier nuclei makes the problem of defining the actual galactic spectrum from these results one of increased difficulty. (2) *The lack of results of high statistical accuracy*: The differential spectra determined to date for the heavier charges have been based on 100 to 200 counts or less for each charge group. To define usefully the differences in spectra of the various charge groups, if they exist, requires counting rates of at least one order of magnitude greater.

As a final comparison of our results with previous emulsion results, the actual charge distributions obtained are shown in Table 4, together with the summary of Waddington[13] for energies >1.5 Bev/nucleon and the recent results of other workers. The charge composition determined by counter techniques agrees quite well with the average of the emulsion results, with perhaps a greater abundance of C and O as recorded in our experiments. All results show a sufficiently wide scatter, however, to indicate that a measuring uncertainty of at least 1 charge unit in the range $4 \leqq Z \leqq 8$ must occur in some of the experiments.

CONCLUSIONS

From the results we conclude the following:

1. Cerenkov-scintillation detector combinations can be used to measure the rigidity and charge

spectra of the heavier nuclei. The energy and charge resolution is comparable with that obtained by the latest emulsion techniques. Use of advanced counter systems can extend the measurements to the lowest energies without serious background corrections. The rapid processing of data and the high counting rate of the counter systems can be used to advantage in investigating characteristics of the spectra related to the origin of the radiation and also to the solar modulation effects.

2. The solar modulation of intensity appears to be identical for all positively charged particles of the same rigidity. We have specifically extended this conclusion to include particles heavier than α particles in the three flights reported in this paper.

3. The measurements near sunspot minimum establish a high degree of similarity in the differential spectra of the various charge components only when the spectra are expressed in terms of the rigidity of the particles. This similarity is sufficient to make it unlikely that the particles have traveled through more than 4 gm/cm^2 of interstellar hydrogen. It appears, furthermore, that the peaks in the differential spectra of the heavier particles that are observed at times near sunspot minimum cannot be due to some residual solar modulation effect or to the effects of ionization loss alone, or ionization loss competing with Fermi acceleration processes in interstellar space. If the form of the differential spectra and their similar rigidity dependence are to be taken seriously, they may reflect the action of large-scale regular fields either near the source region itself or in interstellar space.

4. The relative and absolute intensities of the various charges as measured in this experiment are comparable with those obtained using emulsions. In particular an L/M ratio of 0.36 ± 0.05 essentially independent of rigidity from 2 to 6 Bv is obtained. Such a ratio is consistent with the passage of the primary radiation through 4 to 6 gm/cm^2 of interstellar hydrogen, this value also being independent of rigidity. The amount of hydrogen and, indeed, the method of "passage" through interstellar space depend so critically on actual interstellar conditions as well as on the distribution of sources, however, that such a value is regarded as having definitive meaning only when combined with simultaneous measurements of a number of other characteristics of the heavier nuclei and properties of the cosmic radiation in general.

REFERENCES

1. Fichtel, C. E., "The Heavy Component of the Primary Cosmic Radiation during Solar Maximum," *Nuovo Cimento*, **19**(6): 1100–1115, March 16, 1961.
2. Evans, D., "Observations on the Heavy Nuclei of the Cosmic Radiation at Very Low Energies," in: *Proc. Moscow Cosmic Ray Conf.*, Moscow, 1960, Vol. 3, pp. 92–95.
3. Aizu, H., and Fujimoto, Y., et al., "The Primary Cosmic Radiation at Prince Albert, Canada," in: *Proc. Moscow Cosmic Ray Conf., Moscow*, 1960, Vol. 3, pp. 96–101.
4. ———, and ———, "Heavy Nuclei in the Primary Cosmic Radiation at Prince Albert, Canada. II," *Phys. Rev.*, **121**(4): 1206–1218, February 15, 1961.
5. Tamai, E., "Heavy Nuclei in the Primary Cosmic Rays over Minnesota," *Phys. Rev.*, **117**(5): 1345–1351, March 1, 1960.
6. McDonald, F. B., and Webber, W. R., "Proton Component of the Primary Cosmic Radiation," *Phys. Rev.*, **115**(1): 194–205, July 1, 1959.
7. ———, and ———, "The Variation of the Cosmic Ray Intensity during a Solar Cycle," in: *Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January* 1960, ed. by H. K. Bijl, Amsterdam: North-Holland Publ. Co., 1960, pp. 968–981.
8. ———, "Primary Cosmic-Ray Intensity near Solar Maximum," *Phys. Rev.*, **116**(2): 462–463, October 15, 1959.
9. ———, "Direct Determination of Primary Cosmic-Ray Alpha-Particle Energy Spectrum by New Method," *Phys. Rev.*, **104**(6): 1723–1729, December 15, 1956.
10. Webber, W. R., "Time Variations of Low Rigidity Cosmic Rays During the Recent Sunspot Cycle," in: *Progress in Elementary Particle and Cosmic Ray Physics*, ed. by J. G. Wilson and S. A. Wouthuysen, Amsterdam: North-Holland Publ. Co., Vol. 6, pp. 75–243, 1962.
11. Lockwood, J. A., "On the Long-Term Variation in the Cosmic Radiation," *J. Geophys. Res.*, **65**(1): 19–25, January 1960.
12. Quenby, J. J., and Webber, W. R., "Cosmic Ray Cut-off Rigidities and the Earth's Magnetic Field," *Phil. Mag.*, **4**(37): 90–113, January 1959.
13. Waddington, C. J., "Present State of Knowledge of the Composition of the Primary Cosmic Radiation," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961. *III. Cosmic Rays*, Tokyo: Physical Society of Japan, 1962, pp. 63–68.

14. WEBBER, W. R., "New Determination of the Intensities of Primary Cosmic Ray Alpha Particles and Li, Be, B Nuclei at $\lambda = 41.5°$ Using a Cerenkov Detector," *Nuovo Cimento*, **4**(6): 1285–1306, December 1, 1956.
15. KAPLON, M. F., NOON, J. H., and RACETTE, G. W., "Abundance of Lithium, Beryllium, and Boron in the Primary Cosmic Radiation," *Phys. Rev.*, **96**(5): 1408–1416, December 1, 1954.
16. VAN HEERDEN, I. J., and JUDEK, B., "The Relative Abundances of Cosmic Ray Nuclei of Charge $Z \geq 3$," *Can. J. Phys.*, **38**(7): 964–967, July 1960.
17. O'DELL, F. W., SHAPIRO, M. M., and STILLER, B., "Relative Abundances of the Heavy Nuclei of the Galactic Cosmic Radiation," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm, Kyoto, September* 1961. *III. Cosmic Rays*, Tokyo: Physical Society of Japan, 1962, pp. 23–29.
18. HAYAKAWA, S., ITO, K., and TERASHIMA, Y., "Origin of Cosmic Rays," *Suppl. Progr. of Theor. Phys.*, No. **6**, 1–92 1958.
19. FICHTEL, C. E., "The Multiply Charged Component of the Primary Cosmic Radiation at a Low Energy Cut-Off during Solar-Maximum," Ph.D. Thesis, Washington University, St. Louis, 1960.
20. GINSBURG, V. L., "The Origin of Cosmic Radiation," in: *Progress in Elementary Particle and Cosmic Ray Physics*, ed. by J. F. Wilson and S. A. Wouthuysen, Amsterdam: North-Holland Publ. Co., 1958, Vol. 4, pp. 337–421.
21. DAVIS, L., Jr., "On the Diffusion of Cosmic Rays in the Galaxy," in: *Proc. Moscow Cosmic Ray Conf.*, Moscow, 1960, Vol. 3, pp. 220–225.
22. MEYER, P., "The Cosmic Ray Alpha-Particle Flux during Sharp Forbush Intensity Decreases," *J. Geophys. Res.*, **65**(12): 3881–3887, December 1960.
23. FREIER, P. S., NEY, E. P., and WADDINGTON, C. J., "Lithium, Beryllium, and Boron in the Primary Cosmic Radiation," *Phys. Rev.*, **113**(3): 921–927, February 1, 1959.
24. YAGODA, H., "Observations on Heavy Primary Cosmic Ray Nuclei Above the Atmosphere," U.S. Air Force, Cambridge Research Center, Geophysical Research Paper No. 60, July 1958.
25. BISWAS, S., and LAVAKARE, P. J., et al., "The Energy Spectrum of Nuclei with Charge $Z \geq 6$ in the Primary Cosmic Radiation," *Nuovo Cimento*, **16**(4): 644–670, May 16, 1960.
26. ——, "Intensity of Heavy Nuclei ($Z \geq 3$) of Cosmic Rays Following a Solar Flare," *Bull. Amer. Phys. Soc.*, **5**(4): 258, April 25, 1960 (Abstract).
27. EHMERT, A., "Electric Field Modulation," in: *Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January* 1960, ed. by H. K. Bijl, Amsterdam: North-Holland Publ. Co., 1960, pp. 1000–1008.
28. HAYAKAWA, S., KOSHIBA, M., and TERASHIMA, Y., "Interpretation of the Energy Spectra of Heavy Primary Cosmic Rays," in: *Proc. Moscow Cosmic Ray Conf.*, Moscow, 1960, Vol. 3, pp. 181–192.
29. KORCHAK, A. A., and SYROVATSKY, S. I., "On the Composition of Primary Cosmic Rays," in: *Proc. Moscow Cosmic Ray Conf.*, Moscow, 1960, Vol. 3, pp. 211–219.
30. SINGER, S. F., "The Primary Cosmic Radiation and Its Time Variations," in: *Progress in Elementary Particle and Cosmic Ray Physics*, ed. by J. G. Wilson and S. A. Wouthuysen, Amsterdam: North-Holland Publ. Co., 1958, Vol. 4, pp. 203–335.
31. AIZU, H., and FUJIMOTO, Y., et al., "Heavy Nuclei in the Primary Cosmic Radiation at Prince Albert, Canada; I. Carbon, Nitrogen, and Oxygen," *Phys. Rev.*, **116**(2): 436–444, October 15, 1959.

THE SHAPE OF THE GEOMAGNETIC-FIELD SOLAR-WIND BOUNDARY

GILBERT D. MEAD

Goddard Space Flight Center

DAVID B. BEARD†

University of California

The shape of the boundary of the geomagnetic field in a solar wind has been calculated by a self-consistent method in which, to the first order, approximate magnetic field are used to calculate a boundary surface. The electric currents in this boundary produce magnetic fields, which can be calculated once the first surface is known. These are added to the dipole field to give more accurate fields, which are then used to compute a new surface. This iterative procedure converges rapidly to a particular surface which may be tested by finding how close the total fields outside the boundary are to the required value of zero. The result of this stringent test is that the magnetic fields in the plasma outside the fourth surface and within twice the distance to the boundary on the solar side are everywhere less than 1 percent of the geomagnetic dipole field in the absence of a solar wind. This surface has been used to calculate the perturbation of the geomagnetic field by the solar wind and the results of these calculations, plus a number of applications, are given in an accompanying paper.

INTRODUCTION

The geomagnetic dipole has been observed to be immersed in a continuous plasma stream emanating from the sun (Neugebauer and Snyder, 1962; Freeman et al., 1963; Bonetti et al., 1963). Although small magnetic fields are present in the solar plasma (Coleman et al., 1962), they only serve primarily to "tie" different regions of the plasma together. The foremost feature of the solar plasma stream is the stream pressure it exerts on any obstruction it encounters. Such an obstruction is furnished by the geomagnetic field, which holds off the stream and creates a cavity from which the solar plasma is excluded (Chapman and Ferraro, 1931; Dungey, 1958; Cahill and Amazeen, 1963).

The weak interplanetary magnetic field imbedded in the plasma has little effect on the pressure conditions on the solar side of the earth. [The authors believe that this field is compressed against the boundary of the geomagnetic field (Beard, 1963b), and the pressure on the boundary may be essentially independent of the existence of an interplanetary field; in any case, the particle pressure will change by only a relatively constant factor, less than 2.] On the anti-solar side of the earth, however, it becomes the predominant consideration in determining the closure of the cavity. Just how the cavity does close on the anti-solar side is a very difficult problem, whose solution must await a better knowledge of the interplanetary magnetic field and the theory of plasma motion in the presence of trapped magnetic fields. Fortunately, the electric currents which exist on the boundary on the anti-solar side are very weak and relatively distant from the earth's surface, compared with the solar side boundary. Therefore, our understanding of the solar side boundary shape and the distortion of the geomagnetic field does not depend much on whether an interplanetary field is included in

†National Academy of Sciences—National Research Council Senior Resident Research Associate at Goddard Space Flight Center, Summer, 1963.

our considerations. So there appears to be adequate justification for simplifying the problem by assuming the solar wind to be free from magnetic fields and at zero temperature (no lateral motion of the ions perpendicular to the stream velocity occurs).

Several attempts have been made recently to calculate the shape of the boundary under the assumptions of a field-free, zero-temperature plasma specularly reflected from the boundary (Beard, 1960; Beard, 1962; Beard and Jenkins, 1962; Spreiter and Briggs, 1962; Midgley and Davis, 1963). These have all assumed certain approximations which limits their accuracy. The best of these is the solution presented by Midgley and Davis, which will be compared with ours below. Slutz (1963) has obtained a solution to a problem with somewhat different boundary conditions at the interaction surface. Beard (1963a) has recently made a review of all the calculations made to date. Other recent reviews of the solar wind-earth interaction have been written by Chapman (1963, pp. 371, 421), Hines (1963), and Blum (1963).

A new calculation of the boundary of the geomagnetic field, using a self-consistent method, is presented in this paper. The self-consistent method has been introduced previously (Beard, 1962; Beard and Jenkins, 1962) to illustrate the convergence of the particular approximation to the boundary conditions ordinarily used in the calculation. The accuracy and convergence of the self-consistent method has been illustrated for a variety of simple problems by Baker et al. (1963), who have found that the second or third order approximation is indistinguishable from the exact solution of those problems for which analytic solutions exist.

The method consists of using an approximate magnetic field in the differential equations representing the boundary conditions, from which a first-approximation boundary surface may be computed. The surface currents on this first surface may then be integrated over the surface and added to the dipole field to give a better approximation to the magnetic field just inside the boundary. It is found that by the time the third surface is calculated by this method, negligible improvement over the previous surface is obtained. Magnetic fields due to the surface currents computed within the plasma are almost everywhere found to be within a few parts in a thousand of cancelling the geomagnetic dipole field alone, and thus the plasma is essentially field-free, as required by the boundary conditions. Such excellent cancellation of the field within the plasma leads us to believe that the calculation of the geomagnetic field boundary presented here is everywhere accurate to less than 1 percent, within the assumptions made about the model and the interplanetary medium.

METHOD OF CALCULATION OF THE BOUNDARY

Basic Equations

With the assumptions of a zero-temperature, field-free plasma incident perpendicular to the geomagnetic dipole and undergoing specular reflection, the boundary shape may be determined by equating the stream pressure of the ions on one side of the boundary to the magnetic field pressure on the other side (Chapman and Ferraro, 1931; Ferraro, 1952; Spitzer, 1956; Dungey, 1958; Beard, 1960):

$$p = 2nmv^2 \cos^2\psi = B_t^2/8\pi \qquad (1)$$

where n is the ion density in the solar wind, m is the ion mass, v is the stream velocity, ψ is the angle between the outward vector normal to the boundary and the stream velocity vector, and B_t is the total magnetic field intensity just inside the boundary. If $\hat{\mathbf{n}}_s$ represents the unit outward vector normal to the surface, $\cos\psi = \hat{\mathbf{n}}_s \cdot \hat{\mathbf{v}}$, $\mathbf{B}_t$ must be tangential to the surface, and (1) may be rewritten (Davis and Beard, 1962):

$$|\hat{\mathbf{n}}_s \times \mathbf{B}_t| = -(16\pi nmv^2)^{1/2}\hat{\mathbf{n}}_s \cdot \hat{\mathbf{v}} \qquad (2)$$

where $\hat{\mathbf{v}}$ is the unit vector in the direction of the stream velocity. The minus sign is necessary when the square root is taken, since ψ must always be greater than 90° in order that the boundary surface may be everywhere exposed to the stream pressure.

Equations (1) and (2) are equivalent only if the total field $\mathbf{B}_t$ is really tangential to the boundary. The approximation for $\mathbf{B}_t$ used to calculate a first surface is definitely *not* tangential. However, one test of the self-consistent method is to

determine the extent to which the modified $\mathbf{B}_t$ calculated in higher approximations is actually tangential to the surface. The results of this test will be given in the section on results.

The total magnetic field just inside the boundary, $\mathbf{B}_t$, may be considered to be composed of three parts: (1) the geomagnetic dipole field at that point, $\mathbf{B}_g$; (2) a surface current planor field, $\mathbf{B}_p$, equal to the field produced by an infinite plane sheet of current tangential to the surface; and (3) a surface current field resulting from the curvature of the surface, $\mathbf{B}_c$, calculated at a position at the center of the surface currents. Thus the field just inside the surface due to the surface currents is $\mathbf{B}_p+\mathbf{B}_c$, and the total field at that point is

$$\mathbf{B}_t=\mathbf{B}_g+\mathbf{B}_p+\mathbf{B}_c \tag{3}$$

If we assume that the boundary is infinitesimally thin and move to a position just outside the boundary, $\mathbf{B}_g$ and $\mathbf{B}_c$ are unchanged but $\mathbf{B}_p$ changes sign, and

$$\mathbf{B}_{\text{outside}}=\mathbf{B}_g-\mathbf{B}_p+\mathbf{B}_c=0 \tag{4}$$

since it has been assumed that the plasma is field-free. Thus

$$\mathbf{B}_p=\mathbf{B}_g+\mathbf{B}_c \tag{5}$$

$$\mathbf{B}_t=2(\mathbf{B}_g+\mathbf{B}_c) \tag{6}$$

and (2) becomes

$$|\,\hat{\mathbf{n}}_s\times(\mathbf{B}_g+\mathbf{B}_c)\,|=-(4\pi nmv^2)^{1/2}\hat{\mathbf{n}}_s\cdot\hat{\mathbf{v}} \tag{2'}$$

where thus far no approximations have been made.

Calculation of the first-approximation boundary surface

The curvature field $\mathbf{B}_c$ cannot be determined until we have a first surface, but the first surface cannot be obtained without knowing the field. A reasonable first approximation consists of ignoring the curvature field and solving (2′) for $\mathbf{B}_c=0$. This is equivalent to assuming that the field just inside the surface is equal to twice the component of the geomagnetic field tangential to the boundary (Beard, 1960; Spreiter and Briggs, 1962):

$$|\,\mathbf{B}_t\,|=2\,|\,\mathbf{n}_s\times\mathbf{B}_g\,| \tag{7}$$

The coordinate system used in the calculation is illustrated in figure 1. Spherical coordinates are used, but instead of the usual orientation, the polar or z-axis is anti-parallel to the solar plasma stream, the y-axis points north, and the angle ϕ is measured from the magnetic equatorial plane. In this coordinate system

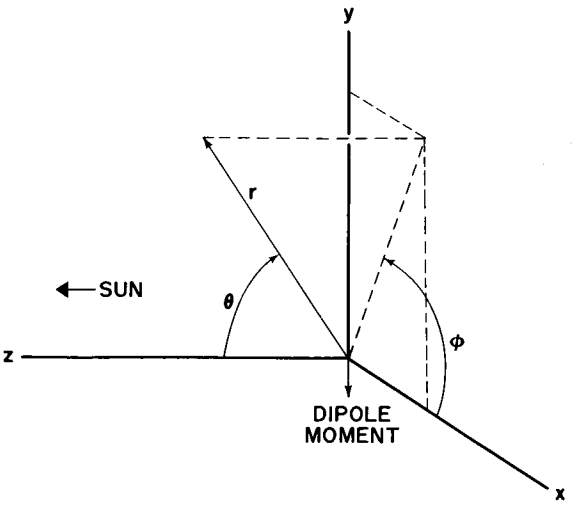


FIGURE 1.—Coordinate system used in the calculation. Note that the polar axis points toward the Sun, instead of the North Star.

$$\hat{\mathbf{n}}_s=a[\mathbf{r}-r^{-1}(\partial R/\partial\theta)\hat{\theta}-(r\sin\theta)^{-1}(\partial R/\partial\phi)\hat{\phi}] \tag{8}$$

$$\hat{\mathbf{v}}=-\cos\theta\mathbf{r}+\sin\theta\hat{\theta} \tag{9}$$

$$\mathbf{B}_g=(M/r^3)(-2\sin\theta\sin\phi\hat{\mathbf{r}}+\cos\theta\sin\phi\hat{\theta}+\cos\phi\hat{\phi}) \tag{10}$$

where the surface is defined by $r=R(\theta,\ \phi)$, $\mathbf{B}_g$ is the geomagnetic dipole field with magnetic moment M, and a is a normalizing factor which makes $\hat{\mathbf{n}}_s$ a unit vector. Since $\hat{\mathbf{n}}_s$ appears on both sides of (2′), this factor cancels out of the equation.

In performing the calculations, all distances are expressed in units of

$$r_0\equiv(M^2/4\pi nmv^2)^{1/6} \tag{11}$$

This is a convenient unit of distance to use, since it represents the altitude of the subsolar point in the first approximation surface, where the field at that point is assumed to be twice the dipole field. When r is expressed in these units,

the constant factor under the square root sign in (2′) is removed, which makes all quantities dimensionless. $\mathbf{B}_c$ is then in units of M/t_0^3.

When (8), (9), and (10) are substituted into (2′), a complicated non-linear, partial differential equation in two unknowns, θ and ϕ, results. A Newton–Raphson technique was used with the aid of an IBM 7094 computer to solve this equation for the entire three-dimensional surface. The surface was divided into a grid of 5° increments in θ and ϕ. It was necessary to determine only a fourth of the entire surface, since the remainder is completely symmetric. The solution starts at the subsolar point, where, because of symmetry, $\partial R/\partial\theta$ and $\partial R/\partial\phi$ are zero. With this simplification, (2′) can be directly solved for r at the subsolar point, giving

$$r_{ss}=(1-B_{c\phi})^{-1/3} \tag{12}$$

in dimensionless units. For the first surface $\mathbf{B}_c=0$, and therefore $r_{ss}=r_0$ as indicated earlier.

In order to obtain the complete surface, the computer moves out from the subsolar point, first along the equator ($\phi=0$) in 5° increments of θ. On the equator $\partial R/\partial\phi=0$ because of symmetry. At each point the computer makes a first guess at a value of r. Then $\partial R/\partial\theta$ is determined by fitting a parabola to this point plus the previous two computed points and finding the slope. These values of r and $\partial R/\partial\theta$ are substituted into (2′), and the difference between the two sides of the equation is computed. In general, this difference will not be precisely zero, since the first guess was not exactly correct. A small variation is made in r, and the new values of r and $\partial R/\partial\theta$ are used to calculate a new difference. If r_1 and r_2 are the two trial values, d_1 and d_2 are the resulting differences, and the variations are sufficiently small that the differences depend linearly upon r in this region, then the "correct" value is

$$r_3=(d_2r_1-d_1r_2)/(d_2-d_1) \tag{13}$$

In practice, r_3 is used as a new guess, another variation is made, smaller than the preceding one, and the process is repeated. The procedure continues until r is obtained to the desired degree of accuracy. In the present case an accuracy of 10^{-7} was required. Since the method is rapidly convergent, this accuracy was achieved with only 2 or 3 variations at each point.

After points on the equator are mapped out, ϕ is increased by 5° and the process continued. Now $\partial R/\partial\phi$ is no longer zero and must also be determined at each point in the same fashion as $\partial R/\partial\theta$ by using the trial value of r plus the points immediately adjacent in ϕ. By this process, the entire three-dimensional surface can be mapped out with great accuracy. The computer takes only about 15 seconds to generate a complete surface. The basic subroutine which keeps track of all the variations was dubbed MAZE, because of the multiple of decisions which it has to make.

Calculation of the curvature field

The next step in the self-consistent method is to use this first surface to compute $\mathbf{B}_c$, the additional magnetic field due to the curvature of the surface (Beard, 1962). This is done by integrating over all the surface currents using the Biot–Savart law:

$$\mathbf{B}_c(\mathbf{r})=\int\frac{\mathbf{J}\times\mathbf{r}'}{|\,\mathbf{r}'\,|^3}\,dS \tag{14}$$

where $\mathbf{J}$ is the surface current per unit length of surface (emu/sec/cm) and $\mathbf{r}'$ is the vector from the differential surface elements dS to the point $\mathbf{r}$, which is taken to be at the center of the surface currents. This curvature field would be identically zero if the surface were a plane.

The direction and magnitude of $\mathbf{J}$ is easily obtained, since the boundary is infinitesimally thin and the current must produce a field discontinuity $B_t=4\pi\mathbf{J}$. Since $\mathbf{B}_t$ is tangential to the surface, (2) and (11) yield

$$\mathbf{J}=\frac{1}{4\pi}\hat{\mathbf{n}}_s\times\mathbf{B}_t=\frac{1}{2\pi}\hat{\mathbf{n}}_s\times(\mathbf{B}_0+\mathbf{B}_c)=\frac{M}{2\pi r_0^3}\cos\psi\hat{\mathbf{j}} \tag{15}$$

where $\hat{\mathbf{j}}$ is a unit vector in the direction of $\hat{\mathbf{n}}_s\times\mathbf{B}_t$. For purposes of machine computation, the expression used for the curvature field was

$$\mathbf{B}_c(\mathbf{r})=\frac{1}{2\pi}\int\frac{[\hat{\mathbf{n}}_s\times(\mathbf{B}_0+\mathbf{B}_c)]\times\mathbf{r}'}{|\,\mathbf{r}'\,|^3}\,dS \tag{16}$$

where the value of $(\mathbf{B}_0+\mathbf{B}_c)$ is the previous approximation value taken at the position of the surface element dS. Thus $\mathbf{B}_c(\mathbf{r}')$ in the integrand

of (16) is assumed zero in computing $\mathbf{B}_c(\mathbf{r})$ for the second order approximation.

The machine calculation of $\mathbf{B}_c$ must be done with care, since a local singularity exists in (16) as $\mathbf{r}' \to 0$. This is a finite singularity, however, since as the surface in the neighborhood of $\mathbf{r}$ approaches a plane, contributions to the curvature field from symmetric surface elements on opposite sides of $\mathbf{r}$ tend to cancel each other. A 5° mesh size was used in the numerical integration, and contributions to the integral were included from every point up to within 5° of θ and ϕ. The region within 5° was excluded. This resulted in an estimated error of less than 5 percent in computing the curvature field. An additional inaccuracy exists in regions near the polar axis of the coordinate system (i.e., near the subsolar point), because the grid in these regions is not very rectangular; therefore the cancellations near the singularity are not as complete. At the subsolar point itself, the symmetry is restored and the value of $\mathbf{B}_c$ calculated here is accurate. The values of $\mathbf{B}_c$ in the region from $\theta=5°$ to $\theta=35°$ were therefore not calculated directly, but quadratic interpolations were made for the three components from $\theta=0°$ to $\theta=40°$, where the grid was reasonably rectangular. Since the calculation of (16) was quite lengthy, computation time was shortened by computing $\mathbf{B}_c$ for every 10° in θ and every 15° in ϕ, and then quadratically interpolating to obtain values at intermediate points. In this manner, the process of determining the curvature field over the whole surface took about 6 minutes on the IBM 7094.

A second surface can now be calculated from (2′) by using the curvature fields from the first surface. A new set of curvature fields is then computed, and this process is continued until the changes from one surface to the next are negligible. One test of the convergence of this method is to see how much $\mathbf{B}_c$ changes from one surface to the next. It was found that the curvature fields calculated from the first surface averaged about 20 percent of the dipole field in the equatorial regions, but were actually larger than the dipole field in many regions near the poles. The curvature fields from the second surface were about 5 percent different, on the average, from those based on the first surface. The third surface fields changed by 1 or 2 percent in the equatorial regions, and about 5 percent near the poles. After this, the changes in the curvature fields were almost everywhere less than 1 percent.

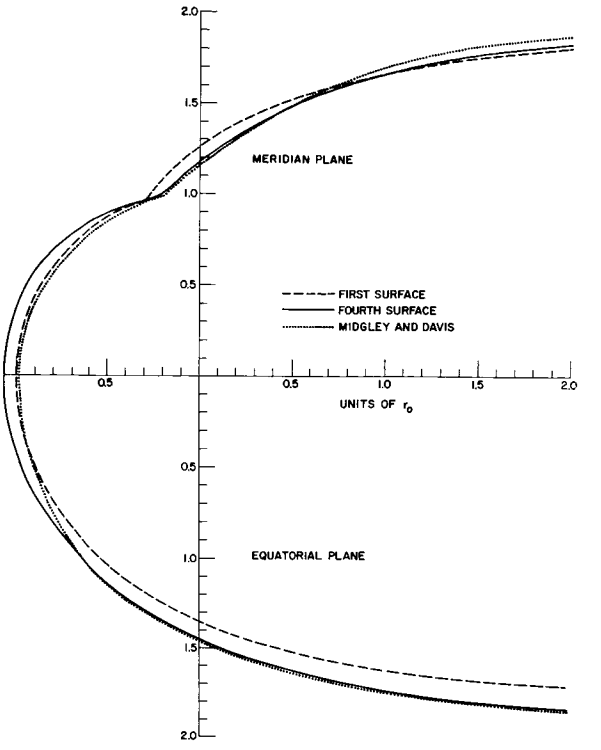


FIGURE 2.—Comparison of the first-approximation surface, the fourth-approximation surface, and the surface obtained by Midgley and Davis (1963), in the meridian and equatorial planes.

RESULTS

The intersection of the first-approximation surface with the equatorial plane and with the noon-midnight meridian plane is shown as a dashed line in figure 2. The trace in the equatorial plane is identical to that calculated by Beard (1960). The trace in the meridian plane is also the same ($r=\text{constant}=r_0$) to within about 20° of the pole on the day side. Past that point, it corresponds essentially to the solution given by Beard, in his equations 27 and 38, using the integration constant of Spreiter and Briggs (1962). The Newton–Raphson method could not give a complete first surface in the meridian plane ($\phi=90°$),

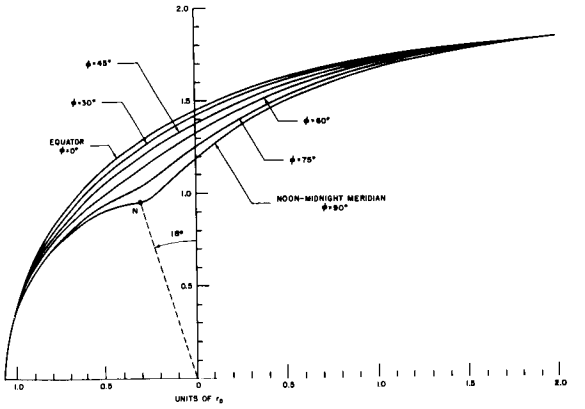


FIGURE 3.—Intersection of the fourth-approximation surface with planes of constant ϕ. The surface at $\phi=15°$ is not shown, since it is almost indistinguishable from the equatorial surface. N is the position of the null point in the noon meridian.

where $\partial R/\partial\phi=0$ because of symmetry. The surface shown is actually a smooth continuation of the surface determined at adjacent values of $\phi=80°$ and $\phi=85°$, which the computer had no difficulty in calculating. In higher orders, when the curvature fields are used in the meridian plane, the difficulty in determining a surface is no longer present, although the shape of the surface changes rapidly in the vicinity of the null point.

A second surface is generated by calculating the curvature fields from the first surface and using these in (2′). This surface is about 7 percent farther out everywhere in the equatorial plane. In the meridian plane, the second surface is outside the first surface near the equator, inside near the pole, and slightly outside again in the distant antisolar regions.

The third and fourth surfaces differ from the second by less than 0.5 percent everywhere near the equator, and therefore are graphically indistinguishable from the second surface in this region. In the polar regions the differences are somewhat greater, the fourth surface differing from the third by as much as 1 percent. A fifth surface has been calculated and it is intermediate between the third and the fourth. Every indication suggests that the process converges very rapidly, and that any number of additional surfaces could be generated with essentially no further change. The process was not carried further than a fifth surface, because the change in the curvature fields from surface to surface was already much less than the accuracy with which these fields are calculated.

The effect of the higher-order approximations can be seen in figure 2, which shows the fourth surface plotted along with the first surface in the two planes of symmetry, $\phi=0°$ (equator) and $\phi=90°$ (meridian). Figure 3 shows the complete three-dimensional fourth surface, giving the shape at intermediate values of ϕ. Figure 4 shows the cross-section of the surface at constant values of θ, i.e., the approximate intersection of the surface with planes perpendicular to the earth-sun line. Here the behavior near the null point can be clearly seen. At very large distances on the dark side, $\theta\approx180°$, the cross-section is almost circular, with a meridional diameter of $3.88r_0$ and an equatorial diameter of $3.78r_0$. Thus, if the solar wind were truly field-free and at zero temperature, and if the boundary were at 10 earth radii at the subsolar point, the diameter of the approximately cylindrical surface cross-section at large distances on the dark side would be about 36 earth radii.

The null point in the meridian plane is at a

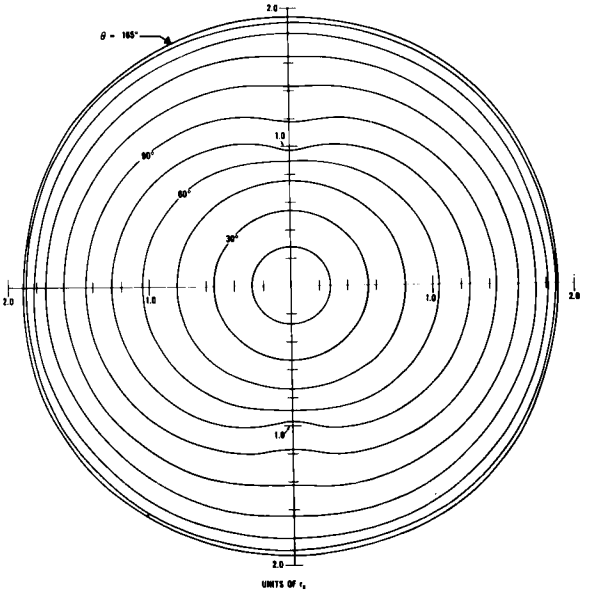


FIGURE 4.—A view of the magnetopause from the position of the sun. Each curve represents a cross section of the surface at constant θ. The surface becomes almost cylindrical in the distant antisolar regions.

TABLE 1.—*Coordinates of the Surface in Units of r_0 (15-Degree Intervals). $\phi=0°$ represents the Equator and $\phi=90°$ is the Noon-Midnight Meridian*

θ	ϕ						
	0	15	30	45	60	75	90
0	1.068	1.068	1.068	1.068	1.068	1.068	1.068
15	1.076	1.075	1.074	1.071	1.069	1.068	1.067
30	1.099	1.096	1.090	1.081	1.072	1.065	1.062
45	1.142	1.136	1.123	1.103	1.079	1.059	1.051
60	1.209	1.201	1.180	1.146	1.102	1.054	1.028
75	1.307	1.297	1.272	1.229	1.168	1.088	1.001
90	1.450	1.440	1.416	1.375	1.319	1.245	1.172
105	1.663	1.656	1.638	1.610	1.571	1.522	1.479
120	1.998	1.994	1.989	1.978	1.959	1.928	1.898
135	2.574	2.576	2.586	2.595	2.594	2.574	2.549
150	3.747	3.757	3.788	3.820	3.838	3.824	3.795
165	7.304	7.330	7.405	7.482	7.529	7.512	7.461

geomagnetic latitude of 72°, i.e., 18° from the pole, on the sunward side. The coordinates of the fourth surface are given in Tables 1 and 2.

In order to better visualize the complete surface, a model was constructed. It is illustrated in figure 5. Templates of the surface for every 15° interval in ϕ were used to cut out aluminum sections, and these sections were joined together in the manner shown. A small sphere with radius $r_0/10$ was placed at the position of the earth. A few of the calculated field lines in the noon-midnight meridian are also shown.

A very stringent test of any proposed plasma-magnetic field boundary has been given by Midgley and Davis (1963). This test is to compute the magnetic field at various points in the plasma as the sum of the field due to surface currents plus the dipole field, and compare this field with the undisturbed dipole field at that point. We have computed the field at $r=1.5r_0$ and $2.0r_0$ in intervals of 15° everywhere on the solar side of the surface. We find that *the ratio of the computed total field in the plasma to the dipole field at every point is less than 1 percent*, and in most cases is a few tenths of a percent. Further out from the surface the surface current field falls off as $1/r^3$, and at $r=10r_0$ the total field is everywhere less than 3 percent of the

TABLE 2.—*Coordinates of the Surface Near the Null Point (5-Degree Intervals)*

θ	ϕ				
	70	75	80	85	90
60	1.069	1.054	1.041	1.031	1.028
65	1.076	1.057	1.037	1.021	1.016
70	1.091	1.066	1.039	1.012	1.003
75	1.116	1.088	1.054	1.014	1.001
80	1.154	1.125	1.090	1.046	1.032
85	1.206	1.178	1.145	1.105	1.091
90	1.271	1.245	1.217	1.183	1.172

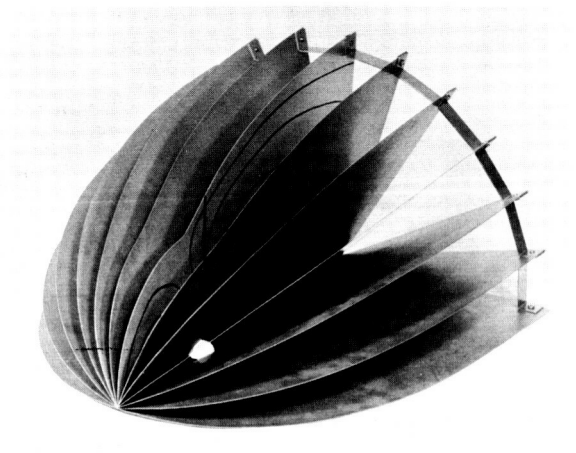


FIGURE 5.—A model of the magnetopause, based on the fourth-approximation surface. Some of the field lines in the noon-midnight meridian plane are shown.

dipole. It is believed that better accuracy could be obtained by improving the calculation of $\mathbf{B}_c$, but this is not justified in view of the changes introduced by the presence of a weak interplanetary magnetic field and time fluctuations in the solar wind pressure.

Another test which can be made is to determine the extent to which the total computed field just inside the boundary, $\mathbf{B}_t$, is tangential to the surface at every point. This was done in two ways. First, $|\ \hat{\mathbf{n}}_s \times \mathbf{B}_t\ | / |\ B_t\ |$, i.e., the ratio of the component of the field tangential to the surface to the total field, was calculated at each point. This ratio must be unity if (1) and (2) are really equivalent. Over most of the surface this ratio was greater than 99.9 percent; it was greater than 99 percent everywhere except within 20° of the null point and within 30° of the anti-solar point. It was greater than 90 percent to within 10° of the null point and the anti-solar point. At these two points, of course, the total field is zero and the ratio is indeterminate.

Secondly, the final surface was recomputed using $|\ \mathbf{B}_t\ |$ instead of $|\ \hat{\mathbf{n}}_s \times \mathbf{B}_t\ |$ to represent the magnetic pressure. The surfaces differed by less than 1 part in a thousand except near the null point and the antisolar point, where changes of as much as 1 or 2 percent were noted.

The fourth surface was compared with the calculation recently published by Midgley and Davis (1963). In order to plot on the same scale, their basic unit of length, R_N, must be compared with r_0. From their equation 5.3,

$$R_N{}^3 = -(M/S_{11}{}^1)(\pi/mnv^2)^{1/2} \tag{17}$$

with $S_{11}{}^1 = -7.0030$, whereas from our (11),

$$r_0{}^3 = (M/2\pi)(\pi/mnv^2)^{1/2} \tag{18}$$

so that

$$R_N = (2\pi/7.0030)^{1/2} r_0 = 0.9645 r_0 \tag{19}$$

The tabular values for Midgley and Davis' "smoothed" surface was multiplied by the above factor and plotted in figure 3 as a dotted line. It is seen that their surface is actually slightly *inside* even the first-approximation surface in the region of the subsolar point, as they have also noted. The fraction of the field just inside the surface contributed by the dipole field can be obtained from (1) and (11). At the subsolar point

$$B_t = (16\pi nmv^2)^{1/2} = 2M/r_0{}^3 \tag{20}$$

and thus

$$\frac{B_g}{B_t} = \frac{M/r_{ss}{}^3}{2M/r_0{}^3} = \frac{1}{2(r_{ss}/r_0)^3} \tag{21}$$

From Midgley and Davis' surface, $r_{ss} = 0.981 r_0$, and therefore 53 percent of the field is contributed by the dipole. For our surface, $r_{ss} = 1.068 r_0$ and 41 percent of the field is contributed by the dipole. With a plane or spherical boundary, the exact percentages are 50 and 35 percent, respectively. The authors feel certain that the dipole must produce less than half of the total field just inside the boundary, and that therefore the true surface must be greater than r_0 at that point. in other words, the curvature field B_c *adds* to the dipole field at the subsolar point rather than opposes it.

Midgley and Davis surface moves out more rapidly than our fourth surface away from the subsolar point, and for all values of θ greater than about 60°, the two surfaces are very close together. Midgley and Davis determine the null point to be 15° from the pole, as opposed to 18° for our surface.

CONCLUSIONS

The surface boundary between the solar wind and the geomagnetic field has been calculated

by use of a self-consistent method in which, in first approximation, the magnetic field resulting from the electric currents on the boundary has been approximated by treating the radius of curvature of the boundary surface as infinite. In succeeding approximations the magnetic field resulting from the currents on the boundary is computed by integration over the preceding surface and used to calculate a new surface. The method converges rapidly, with an ultimate precision limited by the precision with which the magnetic fields are computed. The third and fourth surfaces are almost indistinguishable from the second. The first and fourth approximations have been compared with Midgley and Davis' result and are illustrated in figure 2. The final result, the fourth approximation surface, is shown in more detail in figures 3 and 4. A model of the surface is pictured in figure 5. The accuracy of the fourth approximation surface may be tested by computing the magnetic field outside the boundary (inside the plasma), which should be zero in the absence of an interplanetary magnetic field. The ratio of this computed field to the geomagnetic dipole field out to $r=2r_0$ in the plasma region is everywhere less than 1 percent and of the order of a tenth of a percent in most cases. The computed field just inside the surface used to balance the particle pressure is found to be tangential to the surface except near the null point and in the distant anti-solar regions.

These last tests are such stringent requirements on the solution and are so well satisfied by the final results that we believe that this solution to the theoretical problem (a magnetic dipole in a magnetic-free mono-directional plasma stream) is as accurate a test of the reliability of the calculational method and as accurate a proposed geomagnetic field boundary as could be desired in the light of actual geophysical conditions.

In view of the overwhelming forward pressure of the solar wind (not to mention the time variations in the stream pressure), modifications in the surface calculation introduced by weak interplanetary magnetic fields will be negligible on that part of the surface where the stream pressure is appreciable ($\hat{\mathbf{n}}_s \cdot \hat{\mathbf{v}} \sim 1$, the solar side); moreover, only minor modifications of the magnetic field in the region of the earth computed from these surface currents may be expected in view of the unchanged boundary position on the solar side, where the surface currents are large and close to the earth. Introduction of these higher order effects into the calculations is at present difficult because of our lack of experimental knowledge of the interplanetary magnetic field and lack of theoretical knowledge of the behavior of a high velocity plasma stream containing a trapped magnetic field when encountering an obstacle to its flow. Although pressure balance conditions will remain unchanged on the solar side, the antisolar shape of the boundary will be greatly effected and intuitively may be expected to close in a long rain-drop-shaped tail, rather than remain open as in the proposed problem considered in this paper. These alterations in the position of very weak and very distant surface currents should have only a negligible effect on the distortion of the geomagnetic field of the earth within about 10 earth radii of the surface of the earth. Thus the present calculation furnishes an excellent basis for computation of the magnetic field within the magnetosphere. The result, along with a number of applications, are presented in an accompanying paper (Mead, 1964).

ACKNOWLEDGMENT

The authors would like to thank Mr. Edward Jenkins for this help in the early stages of this calculation.

REFERENCES

BAKER, J. C., BEARD, D. B., and YOUNG, J. C., Self-consistent method for determining the boundary shape between a plasma and a magnetic field, submitted to *Physics of Fluids*, 1963.

BEARD, D. B., The interaction of the terrestrial magnetic field with the solar corpuscular radiation, *J. Geophys. Res.*, **65**, 3559–3668, 1960.

———, The interaction of the terrestrial magnetic field with the solar corpuscular radiation; 2. Second-order approximation, *J. Geophys. Res.*, **67**, 477–483, 1962.

———, The effect of an interplanetary magnetic field on the solar wind, submitted to *J. Geophys. Res.*, 1963b.

———, The solar wind-geomagnetic field boundary, *Rev. Geophys.*, in press, 1963a.

———, and JENKINS, E. B., Correction to the second-approximation calculation of the geomagnetic field, solar wind interface, *J. Geophys. Res.*, **67**, 4895–4896, 1962.

BLUM, R., The interaction between the geomagnetic field and the solar corpuscular radiation, *Icarus*, **1**, 459–488, 1963.

BONETI, A., BRIDGE, H. S., LAZARUS, A. J., ROSSI, B., and SCHERB, F., Explorer 10 plasma measurements, *J. Geophys. Res.*, **68**, 4017–4063, 1963.

CAHILL, L. J., and AMAZEEN, P. G., The boundary of the geomagnetic field, *J. Geophys. Res.*, **68**, 1835–1843, 1963.

CHAPMAN, S., Solar plasma, geomagnetism, and aurora, in *Geophysics: The Earth's Environment, Les Houches 1962 Lectures*, edited by C. DeWitt et al., pp. 371–502, Gordon and Breach, New York, 1963.

———, and FERRARO, V. C. A., A new theory of magnetic storms, Part I: The initial phase, *Terrest. Magnetism Atmospheric Elec.*, **36**, 77–97 and 171–186, 1931.

COLEMAN, P. J., Jr., DAVIS, L., Jr., SMITH, E. J., and SONNETT, C. P., Interplanetary magnetic fields: preliminary Mariner II observations, *Science*, **138**, 1099, 1962.

DAVIS, L., Jr., and BEARD, D. B., A correction to the approximate condition for locating the boundary between a magnetic field and a plasma, *J. Geophys. Res.*, **67**, 4505–4507, 1962.

DUNGEY, J. W., *Cosmic Electrodynamics*, Cambridge University Press, 1958.

FERRARO, V. C. A., On the theory of the first phase of a geomagnetic storm, *J. Geophys. Res.*, **57**, 15–49, 1952.

FREEMAN, J. W., VAN ALLEN, J. A., and CAHILL, L. J., Explorer 12 observations of the magnetospheric boundary and the associated solar plasma on September 13, 1961, *J. Geophys. Res.*, **68**, 2121–2130, 1963.

HINES, C. O., The magnetopause: a new frontier in space, *Science*, **141**, 130–136, 1963.

MEAD, G. D., The deformation of the geomagnetic field by the solar wind, GSFC Document X–640–63.

MIDGLEY, J. E., and DAVIS, L., Jr., Calculation by a moment technique of the perturbation of the geomagnetic field by the solar wind, *J. Geophys. Res.*, **68**, 5111–5123, 1963.

NEUGEBAUER, M., and SNYDER, C. W., Solar plasma experiment; preliminary Mariner II observations, *Science*, **138**, 1095–1097, 1962.

SLUTZ, R. J., Numerical Method for calculating the equilibrium configuration of a free-surface plasma, *Bull. Am. Phys. Soc. II*, **8**, 165, 1963.

SPITZER, L., Jr., *Physics of Fully-Ionized Gases*, Interscience, New York, 1956.

SPREITER, J. R., and BRIGGS, B. R., Theoretical determination of the form of the boundary of the solar corpuscular stream produced by interaction with the magnetic dipole field of the earth, *J. Geophys. Res.*, **67**, 37–51, 1962.

HIGH-ENERGY ELECTRONS IN THE RADIATION BELT

M. P. NAKADA

Goddard Space Flight Center

The birth spectrums of high-energy electrons from the decay of energetic neutrons from cosmic ray albedo have been calculated. Assuming that energy loss is the principle loss mechanism for these energetic electrons, an equilibrium flux of electrons has been derived.

INTRODUCTION

Although the high-energy protons in the radiation belt appear to be explainable through the cosmic ray albedo neutron hypothesis [Freden and White, 1962; Lenchek and Singer, 1962], the origin of the electrons in the radiation belt is unsettled. Calculations by Hess and Killeen [1961], Lenchek, Singer, and Wentworth [1961], and Hess and Poirier [1962] indicate that perhaps a portion of these electrons may be due to the decay of slow neutrons from cosmic ray albedo. However, recent studies of time variations and spectrums [O'Brien, Laughlin, Van Allen, and Frank, 1962; Pizzella et al., 1962] seem to indicate the need for acceleration or other injection processes or both.

The presence of very energetic electrons in the outer belt has recently been confirmed [O'Brien, Van Allen, Laughlin, and Frank, 1962; O'Brien, Laughlin, Van Allen, and Frank, 1962]; the electron energies are definitely greater than is possible from the decay of slow neutrons. Although acceleration mechanisms may be required to explain these energetic electrons, the possibility that they may be due to the decay of fast neutrons from cosmic ray albedo is here examined.

NEUTRON DECAY

Hess, Canfield, and Lingenfelter [1961] give an albedo neutron flux at about 100 km that is approximately E_0^{-2} neutrons/cm^3 sec Mev. Lenchek and Singer [1962] estimate a neutron flux of $2E_0^{-1.8}$ neutrons/cm^2 sec Mev. E_0 in both expressions is the neutron kinetic energies in Mev.

Both neutron spectrums of form $AE_0^{-B}(r_e/r)^2$ have been used where r_e is the radius of the earth and r is radial distance where the flux is considered.

The number of decays per cm^3 sec Mev is given by

$$AE_0^{-B}(r_e/r)^2(\Lambda/C\beta_0\gamma_0) \qquad (1)$$

where $\Lambda=9\times10^{-4}$/sec is the decay constant for neutrons, β_0 is the ratio of the neutron velocity to the velocity of light c, and $\gamma_0=(1-\beta_0^2)^{-1/2}$ is the time dilation factor.

ELECTRON BIRTH SPECTRUM

The electron spectrum due to the decay of a neutron at rest is given by

$$N(\gamma)=0.614\beta\gamma^2(2.53-\gamma)^2 \quad \text{electrons/unit } \gamma$$

where β is the ratio of the electron velocity to the velocity of light and $\gamma=(1-\beta^2)^{-1/2}$.

For a given neutron and electron energy, the total electron energy is given by a Lorentz transform.

$$W=mc^2\gamma_0\gamma(1+\beta\beta_0\cos\theta_0)$$

where W is the total electron energy, mc^2 is the rest energy of the electron, and θ_0 is the angle of emission of the electron in the rest system of the neutron relative to the neutron flight direction.

Rossi [1952] has shown that if monoenergetic

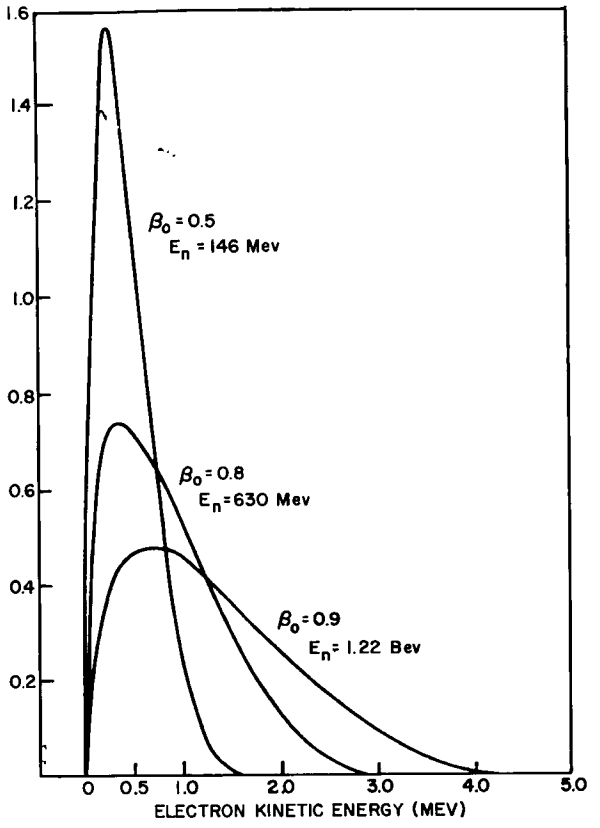


FIGURE 1.—Electron energy spectrum due to the decay in flight of neutrons with various kinetic energies.

particles are emitted isotropically in the rest system, the energy distribution in the laboratory system is constant from $W_{\min}$ to $W_{\max}$ where

$$W_{\min} = mc^2\gamma_0\gamma(1-\beta\beta_0)$$

$$W_{\max} = mc^2\gamma_0\gamma(1+\beta\beta_0)$$

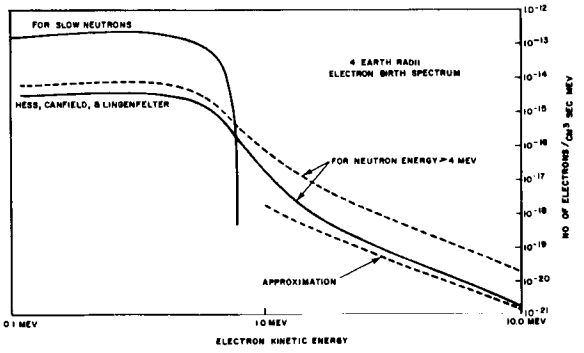


FIGURE 2.—Electron birth spectrum for two fast neutron spectrums and for slow neutrons.

The electron spectrum in the laboratory system has been calculated for neutron energies through the use of this transformation. Figure 1 shows normalized electron spectrums for $\beta_0 = 0.5$, 0.8, and 0.98.

To obtain the electron birth spectrum, small elements of the neutron decay spectrum were transformed through the use of the above calculated laboratory electron spectrum and equation 1. The results are shown in figure 2 for neutron energies greater than 4 Mev.

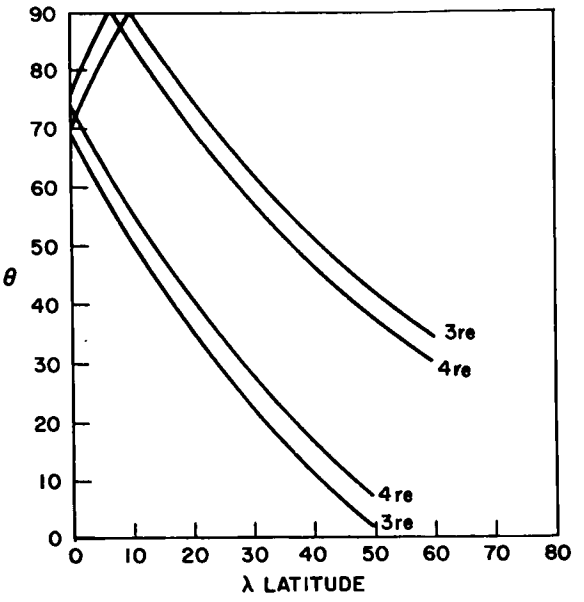


FIGURE 3.—Maximum and minimum angles between fast neutron directions and the magnetic field at 3 and 4 geocentric earth radii.

An approximate result can be obtained by assuming that the average electron is emitted at 90° in the center of mass system having an average $\gamma = 1.6$ (approximately 300 kev). Then $W = mc^2\gamma_0\bar{\gamma}$, and when this is substituted into (1) the electron birth spectrum for $A = 1$ and $B = 2$ becomes

$$S'(W) = \frac{(r_e/r)^2(\Lambda/c)(Mc^2/mc^2\bar{\gamma})}{(Mc^2)^2[(W/mc^2\bar{\gamma})-1]^2[(W/mc_2\bar{\gamma})^2-1]^{1/2}} \cdot (\text{electrons/cm}^3\ \text{sec Mev})$$

where $S'(W)$ is the approximate electron birth

spectrum and Mc^2 is the rest energy of the neutron. $S'(W)$ has been changed to $S'(E)$ where E is the electron kinetic energy; this approximate spectrum is also shown in figure 2.

Fast neutrons in the radiation belt region, in contrast to slow neutrons, travel nearly tangentially from the production layer in the upper atmosphere. Beyond a few thousand kilometers from the earth, neutron directions lie on a cone whose axis passes through the center of the earth. The largest and smallest angles between fast neutron directions and a centered dipole magnetic field line are shown in figure 3 for geocentric distances of 3 and 4 earth radii.

In addition to the pitch angle spread of the electrons introduced by the neutron directions, there is a spread due to the angular distribution of the electrons with respect to the neutron direction. If θ_0 is the center of mass angle of the electron whose velocity is βc, and $\beta_0 c$ is the velocity of the neutron, then the laboratory angle between the electron and neutron direction, θ_L, is given by

$$\tan\theta_L = (\beta \sin\theta_0)/\gamma_0(\beta \cos\theta_0 + \beta_0)$$

Let $\theta_{L1/2}$ be the angle such that one-half the electrons are emitted at smaller angles and one-half the electrons at larger angles. Since electron emission is isotropic in the center of mass system, $\theta_{L1/2}$ corresponds to $\theta_0 = \pi/2$. Then,

$$\tan\theta_{L1/2} = \beta/\gamma_0\beta_0$$

Figure 4 shows $\theta_{L1/2}$ plotted versus neutron kinetic energy for electron energies of 50, 300, and 500 kev in the rest system of the neutron. Since $\theta_{L1/2}$ tends to be quite large, especially for the lower-energy neutrons, the electron source will be assumed to be isotropic. This assumption is poor at higher latitudes and high neutron energies since an appreciably larger fraction of the electrons will be born at small angles to the magnetic field and will be lost in the earth's atmosphere.

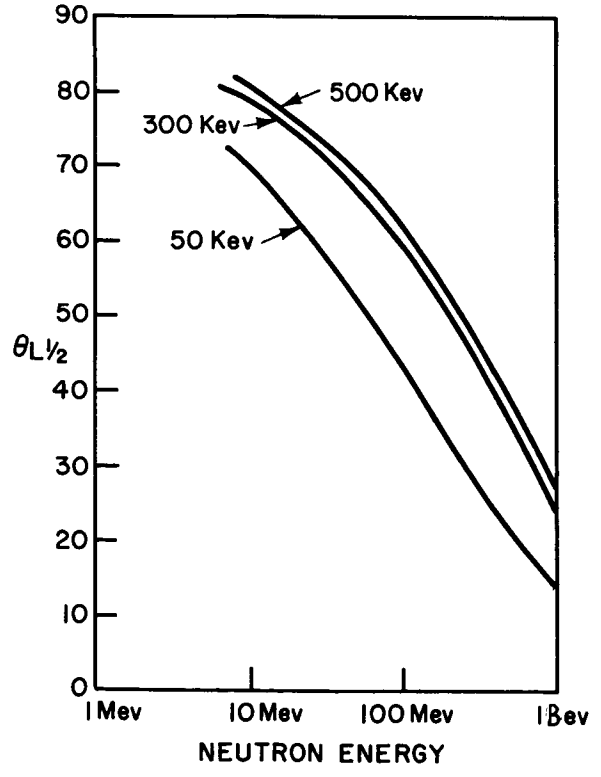


FIGURE 4.—For various laboratory neutron energies and center of mass electron energies, $\theta_{L1/2}$ is the laboratory angle between the neutron direction and the electron such that ½ of the electrons are born at angles less than $\theta_{L1/2}$.

EQUILIBRIUM ELECTRON SPECTRUM AND FLUX

Although the electrons may be lost from the trapping region or lose energy through a number of different processes [see Lenchek, Singer, and Wentworth, 1961], interaction with the ambient atmosphere only will be considered. Hess and Killeen [1961] and Lenchek, Singer, and Wentworth [1961] have indicated that, for the higher-energy electrons, energy loss is a more important loss mechanism than scattering. Energy loss only is considered here; approximate upper limit fluxes are obtained if other loss processes are small.

Lenchek, Singer, and Wentworth [1961] have developed a pure energy loss formula to obtain an equilibrium electron spectrum

$$n(E) = \frac{1}{|\,dE/dt\,|}\int_E^\infty S(E)\,dE \qquad (2)$$

where $n(E)$ is expressed in the same spatial units as the source, $S(E)$, and dE/dt is the collision energy loss:

$$dE/dt = (4\pi r_0^2 c\rho_i/\beta)\,\ln\Lambda_i$$

where $r_0 = e^2/mc^2$, ρ_i is the density of the ith component of the atmosphere, and $\ln\Lambda_i$ is the ratio of the maximum to minimum impact parameter between the electron and the ith component. The main constituent for energy loss in the outer belt will be assumed to be a hydrogen plasma. Spitzer [1956] has tabulated $\ln\Lambda_i$ values for a hydrogen plasma; since $\ln\Lambda$ varies slowly with plasma densities and temperatures, a value of 20 is used.

The equatorial flux has been calculated in two ways. In the first method, the electrons are thought to be in a reflecting container. The equilibrium density is given by (2) where $S(E, a)$ is the number of decays per cm^3 sec Mev at radial distance a. The omnidirectional flux is $J_0 = \beta cn(E, a)$.

In the second method, the contribution of electrons born away from the equator to the equatorial flux is more carefully approximated. Let $S(E, r)$ be the number of electrons born per cm^3 sec Mev at distance r from the center of the earth. Let $S(E, \Omega, a)$ be the equatorial source function at distance a from the center of the earth. Liouville's theorem, the assumed isotropy of $S(E, r)$, and the fact that electron guiding centers oscillate very near a shell formed by revolving a line of force about the magnetic dipole axis are used in deriving $S(E, \Omega, a)$.

Let a_1 be the equatorial distance to a line of force whose equation is $r = a_1 \cos^2\lambda$, where λ is the latitude. The volume enclosed between the equator and some latitude by revolving this line of force about the dipole axis is

$$V_1 = \frac{2\pi}{3}\int_0^{\lambda} a_1^3 \cos^7\lambda \, d\lambda$$

The volume enclosed between two lines of force that are revolved about the dipole axis is

$$V = \frac{2\pi}{3}(a_2^3 - a_1^3)\int_0^{\lambda} \cos^7\lambda \, d\lambda$$

where a_2 and a_1 are the equatorial radii. For a pitch angle α_e at the equator, the contribution of each volume element of V to the equatorial source function is $2S(E, r)/4\pi$. However, only those volume elements that lie between the equator and the mirror latitude for α_e contribute. The mirror latitude is obtained from $\sin^2\alpha_m = 1 = (Bm/B)\sin^2\alpha_e$, where B and Bm are the equatorial and mirror magnetic field:

$$\sin\alpha_e = (\cos^3\lambda m)/(4 - 3\cos^2\lambda m)^{1/4}$$

The equatorial source function per Mev sec ster unit area normal to α_e is given by

$$S(E, \Omega, a) = \frac{a}{\cos\alpha_e}\int_0^{\lambda m} \frac{2S(E, r)}{4\pi} \cos^7\lambda \, d\lambda$$

Along a line of force,

$$S(E, r) = S(E, a)\,(1/\cos^4\lambda)$$

and so

$$S(E, \Omega, a) = \frac{aS(E, a)}{2\pi \cos\alpha_e}\int_0^{\lambda m} \cos^3\lambda \, d\lambda$$

The integral and the approximation to this integral, $0.5\cos\alpha_e$, are shown in figure 5. With this

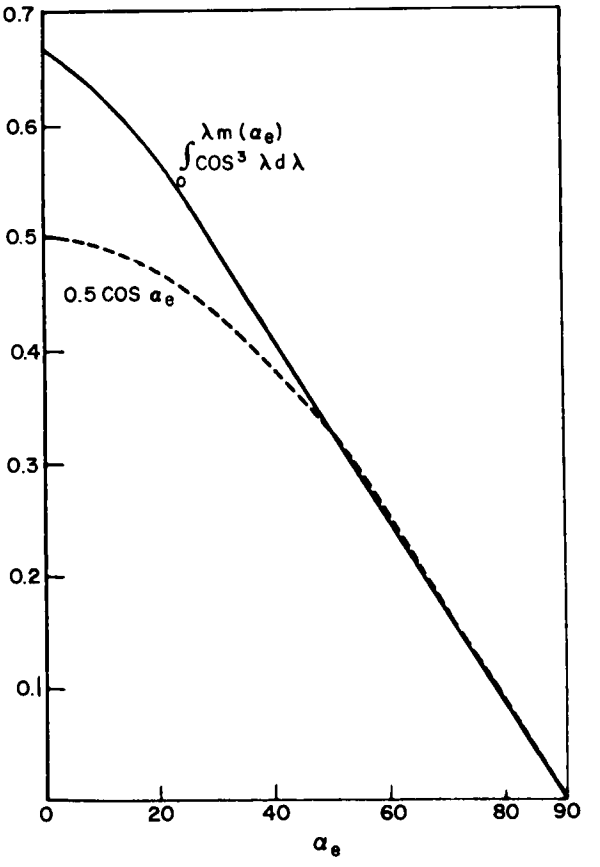


FIGURE 5.—Integral to obtain the source function and the approximation to the integral.

approximation $S(E, \Omega, a)$ becomes

$$S(E, \Omega, a) = aS(E, a)/4\pi$$

The equilibrium electron spectrum is

$$n(E, \Omega, a) = \frac{a}{4\pi \mid dE/dt \mid} \int_E^\infty S(E, a)\, dE$$

The omnidirectional flux is obtained by dividing $n(E, \Omega, a)$ by τ, the travel time from the equator to the mirror point and back to the equator, and multiplying by 4π to obtain contributions from all angles. τ is approximately $0.04a/\beta r_0$ sec with some dependence on α_e. The omnidirectional equatorial flux is [Hamlin, Karplus, Vik, and Watson, 1961]

$$J_0(E, a) = \frac{\beta r_0}{.04 \mid dE/dt \mid} \int_E^\infty S(E, a)\, dE \quad (3)$$

The ratio of the flux obtained by this second method to the flux from the first method is 0.53. The omnidirectional equatorial flux has been evaluated by means of the second method at 4 geocentric earth radii with $\rho_i = 300$ electrons/cm^3. Fluxes due to slow neutron decay and the two fast neutron spectrums are shown in Figure 6. Electrons due to slow neutron decay have previously been calculated by Kellogg [1960], Hess and Killen [1961], Lenchek, Singer, and Wentworth [1961] and Hess and Poirier [1962]. The results in figure 6 are comparable to these and are shown to compare also with electrons from fast neutron decay. Also shown in figure 6 are the measurements by O'Brien, Van Allen, Laughlin, and Frank [1962].

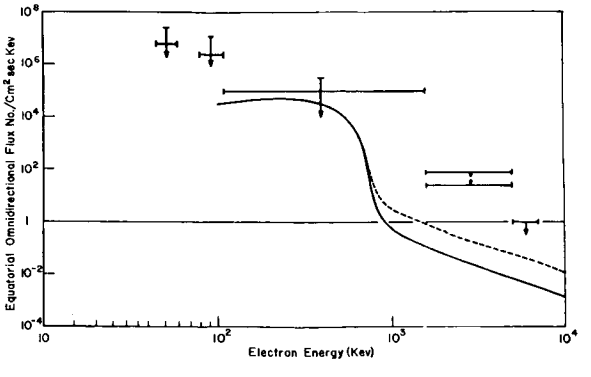


FIGURE 6.—Calculated equilibrium neutron flux assuming energy loss only and using two fast neutron spectrums. Experimental points are also shown.

The contribution of fast neutrons as expected, is mainly above 800 kev and gives approximate power law spectrums.

DISCUSSION

The calculated high-energy electron spectrums depend strongly on the exponent of the power law spectrums for the neutron flux. In these calculations, a difference of 0.2 in the exponent results in a factor of about 5 in the electron flux between 1 and 10 Mev. Although the fast neutron flux is unknown, the agreement between neutron albedo theory and experiment for the high-energy protons in the inner radiation belt give some indication that the values used here are within perhaps an order of magnitude of actual values. The neutron flux above about 1 bev is less certain since few protons at these energies have been observed in the inner radiation belt. However, the absence of these very energetic protons does not necessarily indicate that energetic albedo neutrons do not exist since other loss processes seem to exist for these protons [Lenchek and Singer, 1962]. These same loss processes would probably not apply to the electrons born from these very energetic neutrons since the radii of curvature of the electrons are very small compared to those of protons that are born in the same disintegration.

Experimental measurements that can qualify or make insignificant the fast neutron hypothesis are: (1) the spectrums and absolute flux; (2) variation of the equatorial flux with distance, which should vary approximately as $(\rho r^2)^{-1}$; (3) time variations, since the time for these electrons to lose a Mev of energy is of the order of a hundred years; and (4) equatorial angular distribution.

Three possible explanations for the high-energy electrons ($E > 800$ kev) in the radiation belts are:

a. The fast neutron flux may be sufficient, the atmosphere rare enough, and other loss processes unimportant, so that the fast neutrons may account for all the high-energy electrons.

b. The fast neutron flux may be insufficient to account for all the electrons since energy loss processes are too rapid. However, rather gentle

acceleration processes may exist to compensate for the energy loss. Thus fast neutron decay may inject the energetic electrons [Hess and Poirier, 1962] and an acceleration process may permit a build up to observed fluxes.

c. Energy loss and particle loss from the trapping region may be so large compared with the birth rate that the fast neutrons are inadequate to account for the energetic electrons. However, acceleration processes that act on thermal electrons or more energetic electrons that are injected into the magnetic field from solar emissions exist and produce the required fluxes.

A comparison between these calculations and measurements are for electrons with energy greater than about 1.6 Mev. The neutron spectrums $2E_0^{-1.8}$ and E_0^{-2} give 1.1×10^3 and 150 electrons/cm^3 sec with energy greater than 1.6 Mev. In a recent publication O'Brien, Laughlin, Van Allen, and Frank [1962] found fluxes of the order of 10^3 electrons/cm^2 sec with energy greater than $\sim$1.6 Mev in the outer zone maximum, and this flux did not change appreciably over the period of measurements. However, O'Brien, Van Allen, Laughlin, and Frank [1962] found a flux of about 2×10^5/cm^2 sec for electrons between 1.5 and 5 Mev in the outer radiation zone. This difference may be due to the different locations at which the measurements were made. The results of the first report are from low-altitude and high-latitude observations, whereas the higher flux estimate is from higher-altitude and lower-latitude observations. On the other hand, the difference may be temporal.

Whether the difference is spatial or temporal, fast neutrons from galactic cosmic ray albedo appear to be inadequate to account for the 2×10^5/cm^2 sec flux of energetic electrons. The 2×10^5/cm^2 sec flux for electrons greater than about 1.6 Mev is the interpretation of the counting rate of a 302 Geiger-Müller counter that uses range of materials to determine energy and associated experiments to determine the nature and energy of the particles. Other interpretations of the 302 GM counts and associated experiments are possible. Further experiments may be necessary before fast neutrons are eliminated as a possible source for the energetic electrons in the radiation belt.

NOTE ADDED IN PROOF:

It has been called to the author's attention that similar calculations have been made by Dr. Allen M. Lenchek in his thesis at the University of Maryland.

ACKNOWLEDGMENTS

The numerous folding operations were ably computed by Mr. Dwight Sumida and Mr. William Cahill. Their assistance is gratefully acknowledged. Many discussions with Dr. Wilmot Hess are acknowledged.

REFERENCES

FREDEN, S. C., and WHITE, R. S., Trapped proton and cosmic ray albedo neutron fluxes, *J. Geophys. Res.*, **67,** 25, 1962.

HAMLIN, D. A., KARPLUS, R., VIK, R. C., and WATSON, K. M., Mirror and asimuthal drift frequencies for geomagnetically trapped particles, *J. Geophys. Res.*, **66,** 1, 1961.

HESS, W. N., CANFIELD, E. H., and LINGENFELTER, R. E., Cosmic ray neutron demography, *J. Geophys. Res.*, **66,** 665, 1961.

———, and KILLEEN, J., Spatial distribution of electrons from neutron decay in the outer radiation belt, *J. Geophys. Res.*, **66,** 3671, 1961.

———, and POIRIER, J. A., Energy spectrum of electrons in the outer radiation belt, *J. Geophys. Res.*, **67,** 1699, 1962.

KELLOGG, P. J., Electrons of the Van Allen radiation, *J. Geophys. Res.*, **65,** 2705, 1960.

LENCHEK, A. M., and SINGER, S. F., Geomagnetically trapped protons from cosmic ray albedo neutrons, *J. Geophys. Res.*, **67,** 1263, 1962.

———, ———, and WENTWORTH, R. C., Geomagnetically trapped electrons from cosmic ray albedo neutrons, *J. Geophys. Res.*, **66,** 4027, 1961.

O'BRIEN, B. J., LAUGHLIN, C. D., VAN ALLEN, J. A., and FRANK, L. A., Measurements of the intensity and spectrum of electrons at 1000-kilometer altitude and high latitudes, *J. Geophys. Res.*, **67,** 1209, 1962.

———, ———, ———, and ———, Absolute electron intensities in the heart of the earth's outer radiation zone, *J. Geophys. Res.*, **67,** 397, 1962.

PIZZELLA, G., MCILWAIN, C. E., and VAN ALLEN, J. A., Time variations of intensity in the earth's inner radiation zone, October 1959 through December 1960, *J. Geophys. Res.*, **67,** 1235, 1962.

ROSSI, BRUNO, *High Energy Particles*, Prentice-Hall, Englewood Cliffs, N.J., 1952.

SPITZER, LYMAN, Jr., *Physics of Fully Ionized Gases*, Interscience Publishers, New York, 1956.

SYNCHROTRON RADIATION CALCULATIONS FOR THE ARTIFICIAL RADIATION BELT

M. P. NAKADA

Goddard Space Flight Center

Synchrotron radiation from the artificial radiation belt is calculated in some detail. The angular distribution of the electrons and of the synchrotron radiation is considered. Omnidirectional electron fluxes due to Van Allen, Frank, and O'Brien and Brown and Gabbe are used to derive directional fluxes. Temperatures as a function of frequency and of the geomagnetic latitude at 30 Mc/s are calculated. Results are compared with measurements.

INTRODUCTION

In a recent issue of this journal Peterson and Hower [1963] discussed synchrotron radiation from the artificial radiation belt and presented calculations for sky temperatures that might be expected at the geomagnetic equator according to electron flux estimates by Hess [1963]. In the same issue, different flux estimates by Brown and Gabbe [1963](BG) and Van Allen, Frank, and O'Brien [1963] (VFO) were reported. Synchrotron radiation calculations for the BG and VFO electron fluxes have been made with the inclusion of the angular distributions of the electrons and the synchrotron radiation. The dipole geomagnetic field has been assumed, and geomagnetic latitudes have been used. The fission product electron spectrum due to Carter et al. [1959] has been assumed.

ANGULAR DISTRIBUTION OF ELECTRONS

Figure 1 shows the electron flux maps due to BG and VFO. The dotted parts of the VFO map are extrapolations by the present author. Figure 2 shows his view of equatorial fluxes for BG and VFO; the marked points were taken from those authors.

Given these omnidirectional fluxes, the directional fluxes have been obtained following Ray [1960] and Farley and Sanders [1962]. The omnidirectional flux J along a magnetic field line is plotted as a function of the magnetic field strength. Because of the scarcity of experimental points, analytic forms of J were used to fit the plotted points. Figure 3 compares J values taken from figures 1 and 2 with J values used in the calculations. The analytic form used for BG is proportional to $(B-B_{\max})^2$, and that for VFO is

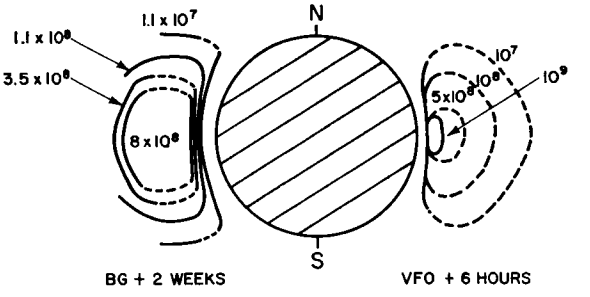


FIGURE 1.—The BG and VFO flux maps used in the calculations. Numbers are omnidirectional fluxes in electrons per centimeter² second.

proportional to $(B-B_{\max})$, where B is the magnetic field for J and $B_{\max}$ is the field where J becomes small. The equatorial omnidirectional flux, J_0, and $B_{\max}$ are then sufficient to determine J along a field line. Equation 6 from Farley and Sanders [1962] has been used to determine the equatorial directional flux from the omnidirectional flux. Liouville's theorem and the equation of motion of the electrons, $\sin^2\theta = (B/B_0)\sin^2\theta_0$,

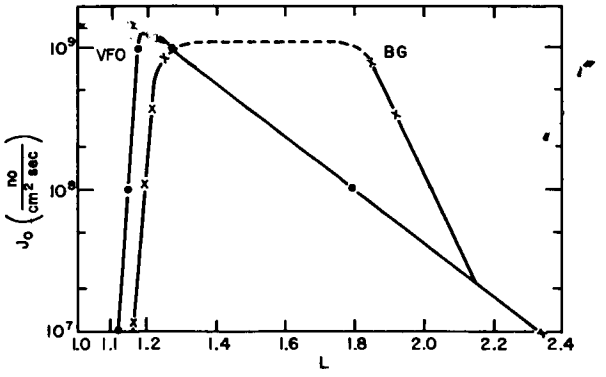


FIGURE 2.—Omnidirectional fluxes along the geomagnetic equator.

have been used to give directional fluxes away from the equator. θ and B are pitch angles and magnetic fields; the subscript zero refers to equatorial values. Figure 4 shows shapes of equatorial directional fluxes as a function of pitch angle.

For later comparison with the width of synchrotron radiation, the half-angle of electron angular distributions at the equator has been plotted in figure 5. The half-angle is the pitch angle measured from the normal to the field line to where the directional flux is one-half of the direction flux at 90° to the field. Figure 6 shows variations in equatorial angles with geomagnetic latitude.

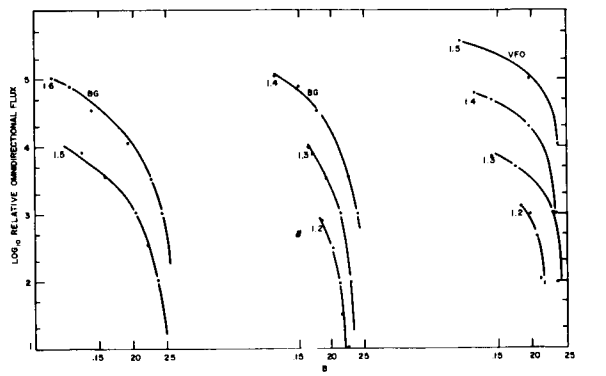


FIGURE 3.—Comparison of relative omnidirectional fluxes vs. B along a magnetic field line. Dots are experimental values. Lines are the assumed analytic forms used in the calculations.

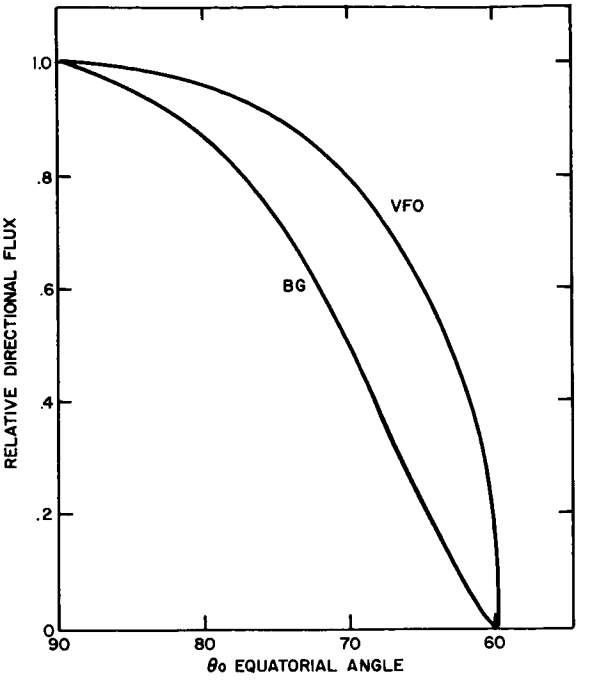


FIGURE 4.—Relative shapes of directional fluxes for the BG and VFO flux estimates.

ANGULAR DISTRIBUTION OF SYNCHROTRON RADIATION

The power radiated in the nth harmonic of the cyclotron frequency into a unit solid angle at angle ψ with the instantaneous orbital plane has been given by Schwinger [1949]. The ψ-depend-

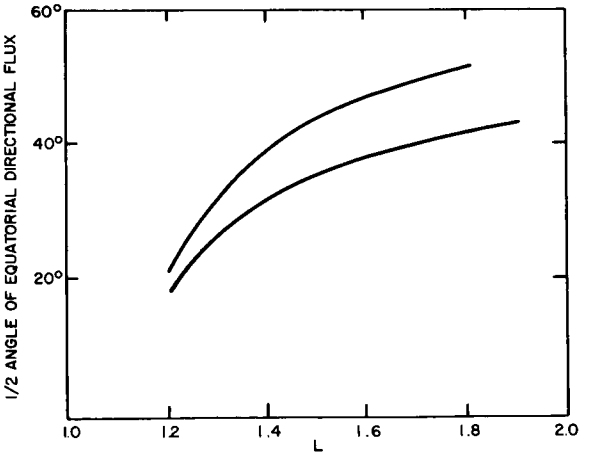


FIGURE 5.—Half-width of equatorial directional fluxes.

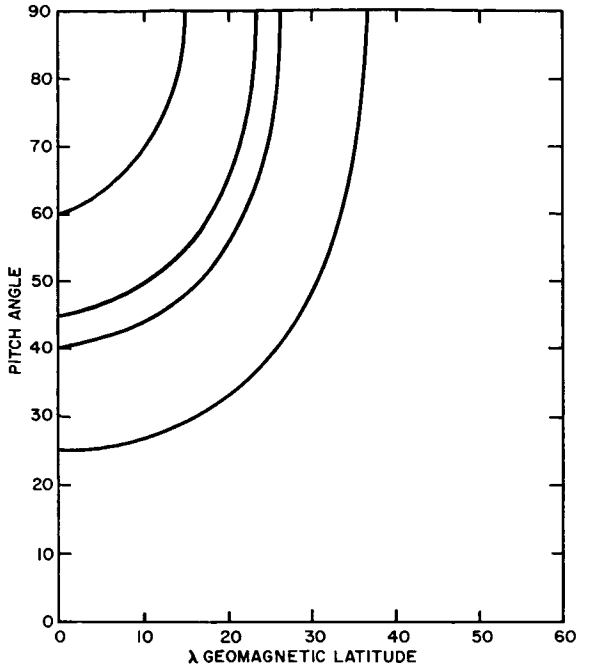


FIGURE 6.—Variation of pitch angle with geomagnetic latitude.

ent part is given by

$$[J_n'(n\beta\cos\psi)]^2+[(\tan^2\psi)/\beta^2]\cdot[J_n(n\beta\cos\psi)]^2$$

where J_n' and J_n are Bessel functions and β is the ratio of the velocity of the electron to the velocity of light. The angular distribution relative to $\psi=0$ has been evaluated. Angles at which the power radiated is one-half of the $\psi=0$ values are plotted in figure 7 for a number of electron energies. The abscissa is $n/(1-\beta^2)^{1/2}$, which is equal

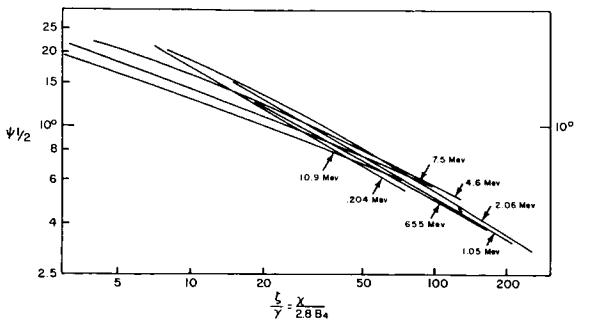


FIGURE 7.—Half-width of the synchrotron radiation pattern. The abscissa is $\nu/(2.8B\sin\theta)$, where ν is the frequency in megacycles per second, B is the magnetic field in gauss, and θ is the pitch angle of an electron.

to $\nu/(2.8B\sin\theta)$, where ν is the frequency in megacycles per second, B is the magnetic field in gauss, and θ is the pitch angle of the electron relative to B. These results show a rather small dependence of the angular distribution on electron energy for a given $\nu/(B\sin\theta)$. At 30 Mc/s for $B=0.2$ gauss, a value that is near the earth, $\psi_{1/2}$ is about 7.8°. At 50 Mc/s for the same B, $\psi_{1/2}$ is about 6.0°. For smaller $B\sin\theta$, $\psi_{1/2}$ is, of course, narrower.

Thus, the angular distribution of the synchrotron radiation is usually narrower than that of the electron angular distribution for frequencies and magnetic fields of interest in this report.

CALCULATION OF TEMPERATURES

The brightness along a line of sight is given by

$$\int \frac{dP}{d\Omega}\,dR$$

where $dP/d\Omega$ is the power emitted per unit volume along the line of sight per steradian. The integration is along the line of sight. Since the synchrotron radiation pattern is usually narrower than that of the electrons, the following approximation has been used for the brightness b:

$$b(\nu)=\int P(\nu, B\sin\theta)[j(\theta)/\beta c]\,dR \qquad (1)$$

$P(\nu, B\sin\theta)$ is the total power emitted per unit frequency at frequency ν for an electron moving toward the observer. This electron has a pitch angle θ with the magnetic field B. $j(\theta)$ is the directional flux of fission product decay electrons. $P(\nu, B\sin\theta)$ is given by

$$P(\nu, B\sin\theta)=\int P(E, \nu/[B\sin\theta])N(E)\,dE$$

where E is the electron energy, $N(E)$ is the normalized electron spectrum, and $P(E, \nu/B\sin\theta)$ is the total power emitted per unit frequency at ν and $B\sin\theta$ for an electron with energy E.

Although equation 1 is an approximation, errors introduced through its use are small for electron and radiation pattern widths considered in this report. To obtain an exact $dP/d\Omega$, the radiation pattern (slightly different at each angle and electron energy) should be folded into the

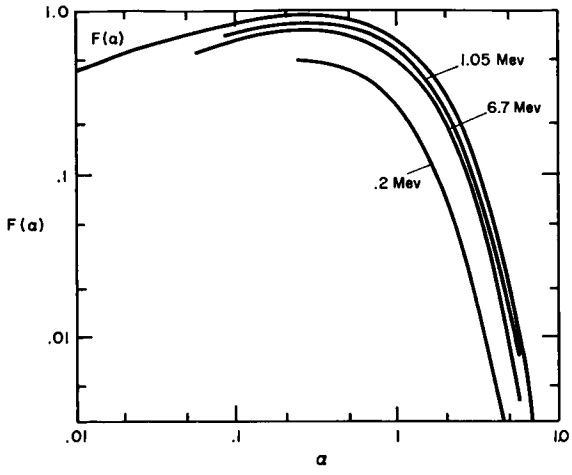


FIGURE 8.—Comparison between relativistic and nonrelativistic synchrotron radiation-power formula.

electron angular distribution, and the result of this fold along the line of sight should be used. If the electron and radiation distributions are assumed to be Gaussian, the result of the fold will give a pattern whose width is given by the square root of the sum of the squares of the assumed Gaussians. For most distributions considered in this report, the resulting Gaussian is only a few per cent broader than an appropriate electron distribution Gaussian. Since both the electron and radiation distributions fall off more rapidly than appropriate Gaussians, the true fold should be narrower than the result of folding the Gaussians. To assess the effect of folded patterns that are broader than the electron pattern, calculations have been made by increasing the electron width to exceed the result of the fold at all points. The calculated temperatures indicate that errors introduced through the use of equation 1 are small.

The relativistic formula for $P(\nu, B\sin\theta, E)$ has been used [Westfold, 1959]

$$P(\nu, B\sin\theta, E) = CB\sin\theta\alpha\int_{\alpha}^{\infty} K_{5/3}(\eta)\, d\eta \quad (2)$$

C is a constant equal to 2.34×10^{-22} erg/sec cycle/sec,

$$\alpha = \frac{\nu(\mathrm{Mc/s})}{4.2\gamma^2 B\sin\theta}$$

$$\gamma = (1-\beta^2)^{-1/2}$$

and $K_{5/3}$ is a Bessel function. This formula gives upper limit values.

$$\alpha\int_{\alpha}^{\infty} K_{5/3}(\eta)\, d\eta$$

is plotted in figure 8. Corresponding values for the nonrelativistic formula [Schwinger, 1949, formula III.28] are also shown in figure 8. These results indicate that, where radiation by electrons is efficient, the relativistic formula is a satisfactory though upper-limit approximation.

Equation 1 with $\beta = 1$ has been used to evaluate sky brightness. The Rayleigh–Jeans approximation to the Planck radiation formula has been used to convert brightness to temperatures, T; $b = 2kT/\lambda^2$, where k is Boltzmann's constant and λ is the wavelength.

Figure 9 shows sky temperatures at 30 Mc/s for the BG flux for an observer at 20°N geomagnetic latitude.

Figure 10 shows temperatures as a function of frequency for the BG and VFO fluxes for an observer at the geomagnetic equator looking

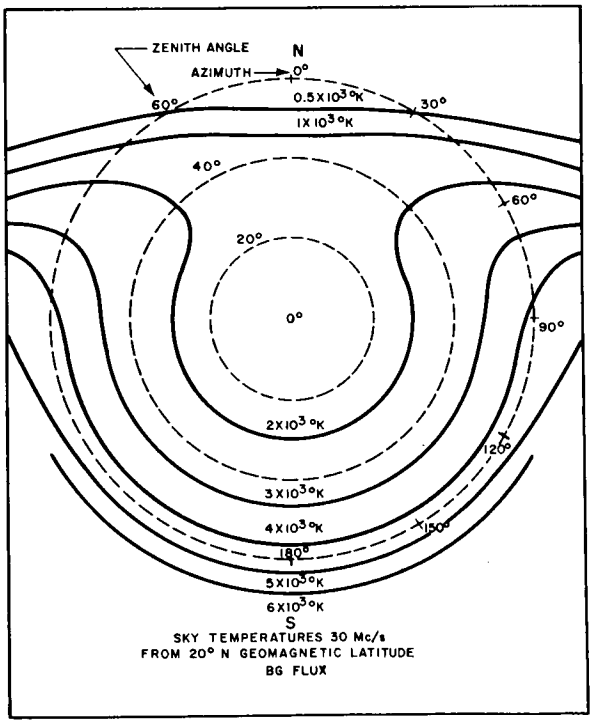


FIGURE 9.—Sky brightness map due to the artificial radiation belt. Solid lines are isophotes.

vertically. At the geomagnetic equator, the temperatures calculated for a narrow-beam antenna and a $\cos^2 Z$ antenna are very nearly the same. Z is the zenith angle.

Despite the time difference for the two flux estimates (VFO at 6 hours and BG at 2 weeks after the detonation), the results are very nearly the same. The VFO values tend to be higher at higher frequencies. This is to be expected, since VFO fluxes are higher at lower altitudes and higher frequencies show a greater decrease in radiation with altitude [Peterson and Hower, 1963].

Experimental points are taken from measurements of Ochs et al. [1963] with times after shot adjacent. Their 0600 local time background temperature at 50 Mc/s of 5200°K was obtained from their sky survey [Ochs et al., 1963]. The 30-Mc/s background temperature of 15,000°K at 0600 local time was obtained from their ratio of 30-Mc/s to 50-Mc/s temperatures before the shot.

Temperatures have also been calculated at 30 and 50 Mc/s for an estimated flux for VFO at 2 weeks after the shot. The BG equatorial flux has been used from $L=1.16$ to $L=1.28$; the VFO flux has been used from $L=1.28$ to $L=2.3$. The BG angular distribution has been used. Temperatures for this flux estimate are labeled 2W-VFO and are also plotted in figure 10.

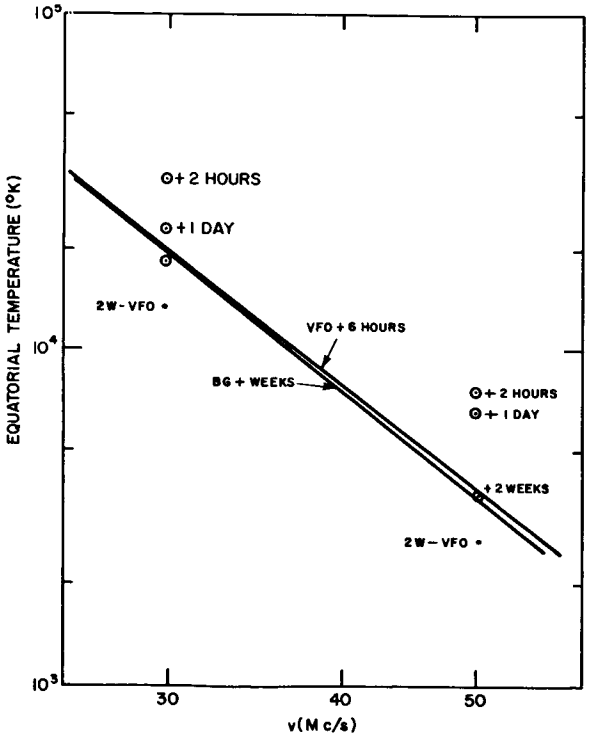


FIGURE 10.—Comparison between calculated and observed sky temperatures at the geomagnetic equator.

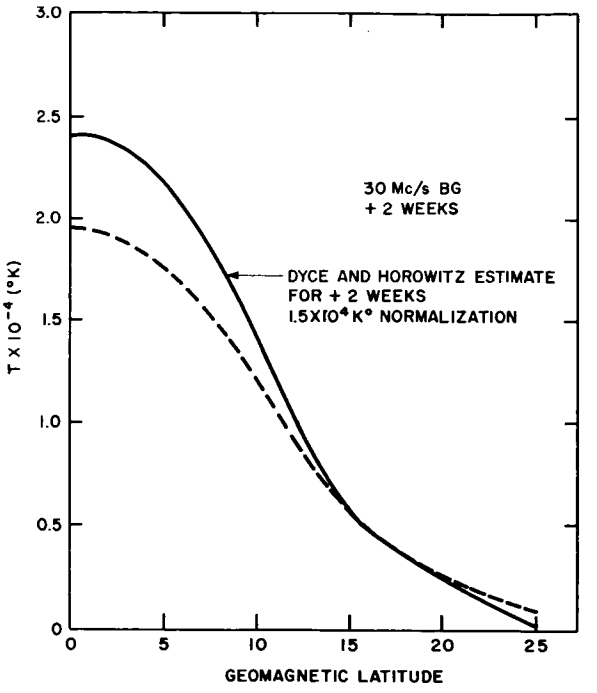


FIGURE 11.—Dotted line shows predicted sky temperatures for vertically directed antennas with $\cos^2 Z$ pattern for the BG flux estimates at +2 weeks at different geomagnetic latitudes. Solid curve is the experimental curve of Dyce and Horowitz; this curve has been renormalized, and a decay has been included.

The calculated changes in temperatures with geomagnetic latitude at 30 Mc/s for observers looking vertically with $\cos^2 Z$ antennas are plotted in figures 11 and 12 for the BG and VFO flux estimates. The experimental measurements from Dyce and Horowitz [1963] are temperatures normalized to preshot diurnal minimum for an approximately $\cos^2 Z$ antenna. Dyce and Horowitz [1963] used 10^4 deg K for this minimum temperature. This seems rather low for their antennas, since interpolation of Turtle et al.'s [1963] results at 26.3 and 38 Mc/s gives a minimum temperature of 10^4 deg K at 30 Mc/s for a 15° right ascension by 44° declination antenna. Estimates for a

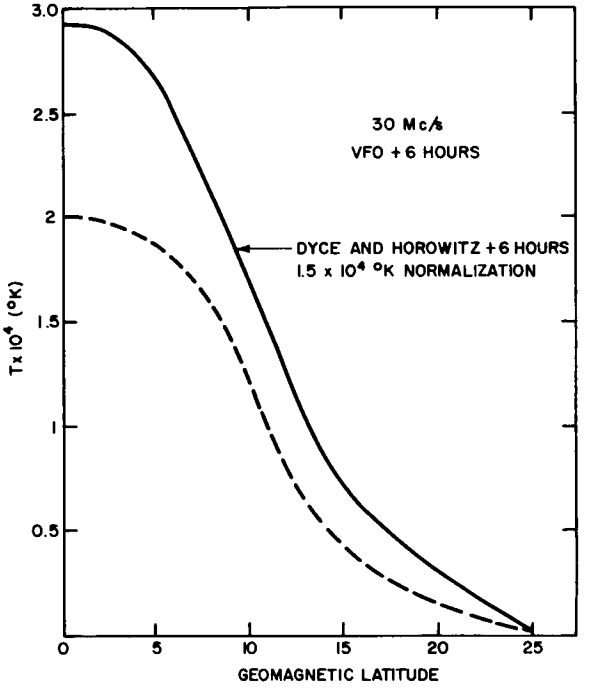


FIGURE 12.—Dotted line shows predicted sky temperatures at +6 hours for vertically directed antennas with $\cos^2 Z$ pattern for the VFO flux estimate at different geomagnetic latitudes. Solid curve is the renormalized experimental curve of Dyce and Horowitz.

$\cos^2 Z$ antenna, using the absolute temperature maps due to Turtle et al. [1963] and Steiger and Warwick [1961], indicate that the normalization temperature at 30 Mc/s for the diurnal minimum was probably near 1.5×10^4 deg K. A latitude distribution using this estimated normalization temperature is plotted in figures 11 and 12. For the BG comparison, a $(1+t/T)^{-1}$ decay with $T=60$ days has been used at all latitudes, although +2 week measurements at the geomagnetic equator are in agreement with calculated temperatures (see figure 10).

DISCUSSION

The results of calculated equatorial temperatures seem to indicate that the VFO flux may be inadequate to explain measured temperatures at +6 hours. The BG flux appears to give satisfactory agreement with observations at 2 weeks.

The latitude distribution is more sensitive to fluxes at larger L values. The VFO flux shows greater temperature deviations from observations at higher latitudes. Whether the BG flux estimate is in agreement with observations will depend on measurements of absolute temperatures for the antennas that were used and on measurements at +2 weeks.

The largest difference between the two flux estimates is for L values greater than 1.28. If a fission spectrum is assumed for the difference between the two flux estimates ($L>1.28$), the calculated equatorial temperature for this difference flux at 50 Mc/s is 941°K. The BG angular distribution has been used in this calculation, which is for a narrow-beam antenna pointed vertically. This temperature may be compared with the minimum detectable temperature of about 40°K at 50 Mc/s [Ochs et al., 1963] using polarization techniques. Ochs et al. [1963] attempted to observe synchrotron radiation from the natural radiation belt immediately before the detonation but detected no positive signal. From this it may be inferred that, if the difference between the two flux estimates was natural, but had a spectrum similar to that of fission product decay electrons, it should have been detectable. Electrons with energy less than 2 Mev do not contribute appreciably to 50-Mc/s radiation for this difference flux.

BG are unable to distinguish any spectral difference between regions where the fission spectrum was highly probable and the difference-flux regions. However, their detector is not sensitive enough to distinguish between spectrums in which the greater-than-2-Mev parts may be very different.

VFO find from their Injun I data that the artificial belt spectrum is considerably 'harder' than the natural one at L of about 4, but whether this hardening extends to equatorial regions in the same way is not known.

Although Explorer 12 traversed this region and had a detector sensitive to electrons with energy greater than 1.6 Mev, detailed studies have not been published. Perhaps a study of Explorer 12 results and spectral and decay studies of Explorer 14 and 15 and Telstar results in conjunction with the negative polarization results of synchrotron radiation before the shot will shed light on what fraction of the BG observations was natural.

CONCLUSIONS

Angular distributions of electrons have been used that are consistent with published omnidirectional flux maps of BG and VFO.

The angular distribution of synchrotron radiation has been found to be considerably narrower than that of electrons for most situations. The fold of the synchrotron radiation pattern onto the electron angular distribution should be slightly wider than the electron pattern.

Since the electron distribution has been used to represent the result of the folded radiation and electron distribution, calculations have been made with electron distributions that were broader than the result of the fold at all locations. The deviations due to this broadening were found to be small.

Although the relativistic radiation formula tends to give upper-limit results, evaluation of the nonrelativistic formula shows that the relativistic approximation is quite adequate even at frequencies lower than those considered here.

For the assumed magnetic field, electron spectrum, and fluxes, these calculations have been estimated to have an uncertainty of about 10 percent.

The VFO flux estimate has been found inadequate to explain observed temperature at the equator and at higher latitudes.

Conclusions about the adequacy of the BG estimate will depend on absolute calibration of antennas and on later results, although present estimates of absolute temperatures for the antennas indicate good agreement between measured and calculated temperatures.

Calculated temperatures for the difference between the BG and VFO fluxes for L greater than 1.28 set limits on electron spectrums in the natural radiation belt before the shot. This result, in conjunction with satellite measurements, may determine what fraction of the BG flux estimate was natural.

ACKNOWLEDGMENTS

I gratefully acknowledge the work of Tom Michels, who made the calculations involving Bessel functions. Robert Baxter set up the code for the detailed temperature calculation; his work is gratefully acknowledged.

REFERENCES

Brown, W. L., and Gabbe, J. D., The electron distribution in the earth's radiation belts during July 1962 as measured by Telstar, *J. Geophys. Res.*, **68,** 607–618, 1963.

Carter, R. E., Reines, F., Wagner, J. J., and Wyman, M. E., Free antineutrino absorption cross section, 2, Expected cross section from measurements of fission fragment electron spectrum, *Phys. Rev.*, **113,** 280, 1959.

Dyce, R. B., and Horowitz, S., Measurement of synchrotron radiation at central Pacific sites, *J. Geophys. Res.*, **68,** 713–721, 1963.

Farley, T. A., and Sanders, N. L., Pitch angle distributions and mirror point densities in the outer radiation zone, *J. Geophys. Res.*, **67,** 2159–2168, 1962.

Hess, W. N., The artificial radiation belt made on July 9, 1962, *J. Geophys. Res.*, **68,** 667–683, 1963.

Ochs, G. R., Farley, D. T., Jr., Bowles, K. L., and Bandyopadhay, P., Observations of synchrotron radio noise at the magnetic equator following the high-altitude nuclear explosion of July 9, 1962, *J. Geophys. Res.*, **68,** 701–711, 1963.

Peterson, A. M., and Hower, G. L., Synchrotron radiation from high-energy electrons, *J. Geophys. Res.*, **68,** 723–734, 1963.

Ray, E. C., On the theory of protons trapped in the earth's magnetic field, *J. Geophys. Res.*, **65,** 1125–1134, 1960.

Schwinger, J., On the classical radiation of accelerated electrons, *Phys. Rev.*, **75,** 1912, 1949.

Steiger, W. R., and Warwick, J. W., Observations of cosmic radio noise at 18 Mc/s in Hawaii, *J. Geophys. Res.*, **66,** 57–66, 1961.

Turtle, A. J., Pugh, J. F., Kenderdine, S., and Pauliny-Toth, I. I. K., The spectrum of the galactic radio emission, 1, Observations of low resolving power, *Monthly Notices Roy. Astron. Soc.*, **124,** 297, 1962.

Van Allen, J. A., Frank, L. A., and O'Brien, B. J., Satellite observations of the artificial radiation belt of July 1962, *J. Geophys. Res.*, **68,** 619–627, 1963.

Westfold, K. C., The polarization of synchrotron radiation, *Astrophys. J.*, **130,** 241, 1959.

VARIATIONS OF THE PROTON ENERGY SPECTRUM WITH POSITION IN THE INNER VAN ALLEN BELT

J. E. NAUGLE AND D. A. KNIFFEN

Goddard Space Flight Center

A cylindrical stack of nuclear emulsions was flown in the payload section of a research rocket into the northern edge of the inner part of the Van Allen radiation region on September 19, 1960. The emulsion stack was shielded inside a tungsten cassette and was then exposed to the ambient radiation by rotation behind a window in the wall of the cassetts covered with a 0.4-g/cm² aluminum shield. In this way it was possible to make a detailed study of the energy spectrum of the trapped particles at various positions in space. The rocket trajectory covered L values ranging from about 1.45 to $1.85R_E$. Five points along the trajectory were chosen for analysis, and the differential and integral proton energy spectrums were obtained for each point. Above 30 Mev in all cases, and below 30 Mev at the lower L values, the proton energy spectrums were in agreement with previous experiments; however, there was a sharp change in the slope of the spectrums with a steeper slope [$j(E)=KE^{-4.5}$] at the higher $L(L>1.6R_E)$ values than had been expected on the basis of previous experiments and the galactic cosmic ray neutron albedo theory. The data for each of the points are presented, and the results are discussed. No particles of mass greater than the proton were observed in a particle composition study averaged over the trajectory. An upper limit is presented for heavier particles in the trapped radiation.

INTRODUCTION

The work of Van Allen et al. [1958] and others demonstrated the existence of the trapped radiation region in the vicinity of the earth and showed that there are large variations in intensity with position. Freden and White [1960, 1962] and Armstrong et al. [1961] have shown the presence of penetrating protons in the inner zone and have measured their energy spectrum averaged over an ICBM trajectory from Cape Canaveral.

The emulsion experiments by Freden and White and by Armstrong et al. were limited in the amount of information that could be obtained, because they were hitchhikes on military vehicles. The lowest energy that could be detected was determined by the thickness of the nose cone. There was no mechanism to determine the flux as a function of position in the belt. The early satellite experiments could accurately measure the flux as a function of position but could not differentiate between electrons and protons or measure their energy. The present experiment was conceived to make use of the nuclear emulsion properties to determine accurately not only the number of particles but also their velocity, mass, and charge as a function of position. For this purpose a special nose cone, the NERV (nuclear emulsion recovery vehicle), was developed, flown, and recovered successfully on the first attempt, September 19, 1960. To measure the variation of the flux over the greatest possible region of the belt, the rocket was flown on a special trajectory from the Pacific Missile Range which carried it along a magnetic meridian into the inner belt. Preliminary results of the experiment have already been published [Naugle and Kniffen, 1961].

The spatial and temporal variations of j are required to map out the radiation levels in the belts adequately as well as to determine the source or sources of the particles, their lifetimes

in the trapped region, the mechanism limiting the intensities of the belts at higher altitudes, and various other features of the belts.

Several authors [Singer, 1958; Vernov et al., 1959; Hess, 1959] have suggested the decay protons of cosmic ray neutron albedo as at least a partial source of the protons in the inner belt. Lenchek and Singer [1962] and Lenchek [1962] have also calculated the effect of a solar cosmic ray neutron albedo source. An abundance of more recent data [Freeman, 1962; Bame et al., 1961; Davis, 1962; Pizella et al., 1962; Imhof et al., 1962] indicates that there must be additional sources of intense numbers of protons down to energies of a few kilo electron volts in both the inner and the outer belts. Both direct injection and local acceleration have been suggested as possible sources, but as yet there is no quantitative explanation of the observed spectrums in this energy region.

Several loss mechanisms have been presented to explain the absence of trapped particles at higher energies at greater distances from the earth. Singer [1959] and Lenchek and Singer [1962] have suggested that the adiabatic invariance of the magnetic moment will break down at a momentum such that the radius of gyration of the proton is comparable to the dimensions of the inhomogeneities in the magnetic field. Welch and Whitaker [1959], Dragt [1961], and Wentzel [1961] have suggested that hydromagnetic waves scatter the more energetic particles magnetically, causing them to mirror lower in a more dense atmosphere and be removed. Other mechanisms, such as an instability of the earth's magnetic field [Gold, 1959], the effect of the large magnetic anomaly at Capetown [Dessler, 1959], and absorption by interstellar dust [Lenchek and Singer, 1962], have also been suggested.

The purpose of this paper is to discuss the results of the NERV experiment more completely and to compare the predictions of the various theories of the origin, acceleration, and loss mechanisms of the trapped protons with these results.

DESCRIPTION OF THE EXPERIMENT

The experiment consisted of a cylindrical stack of 20-Ilford G-5 and G-2 nuclear emulsions. The stack was encased in a tungsten emulsion container with a wall at least 30.6 g/cm² thick everywhere except for a port, 1 cm² in area, on the side of the container. The port was covered by aluminum foil 0.04 g/cm² thick.

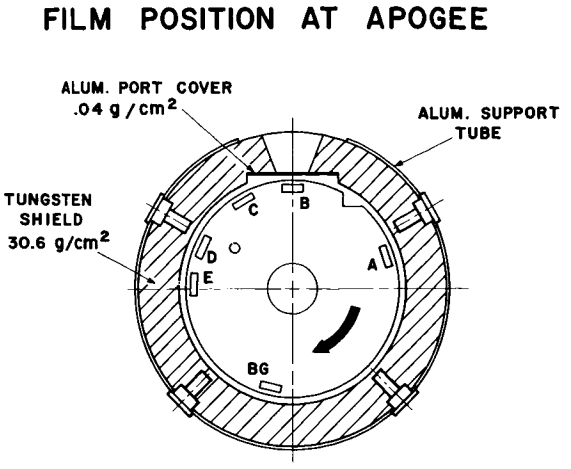


FIGURE 1.—Cutaway top view of the nuclear emulsion container. Film position at apogee.

Figure 1 shows the arrangement. The emulsion stack began rotating at liftoff of the rocket at a constant rate of 0.221°±0.001°/sec. It rotated for 1440 seconds, turning through 318°. Each exposed point on the periphery of the stack was behind the port for 79 seconds, and behind 30.6 g/cm² of tungsten for the remainder of the flight. The regions labeled *A*, *B*, *C*, *D*, and *E* in figure 1 represent areas that were scanned to measure the spectral intensity at a given time during the flight. The track population in these regions is made up of tracks of particles that came through the port during the 79-second exposure and a background of particles that passed through the tungsten during the remaining time the payload was in the radiation belt. The region of the stack labeled *BG*, which did not pass behind the port, has been used to measure this background.

This arrangement improved upon previous emulsion experiments in two respects. The rotating emulsion made it possible to determine the time, and hence the position, at which a given particle had entered the emulsion. The very thin window over the port, together with the arrangement for extending the emulsion container outside the nose cone, lowered the detection thresh-

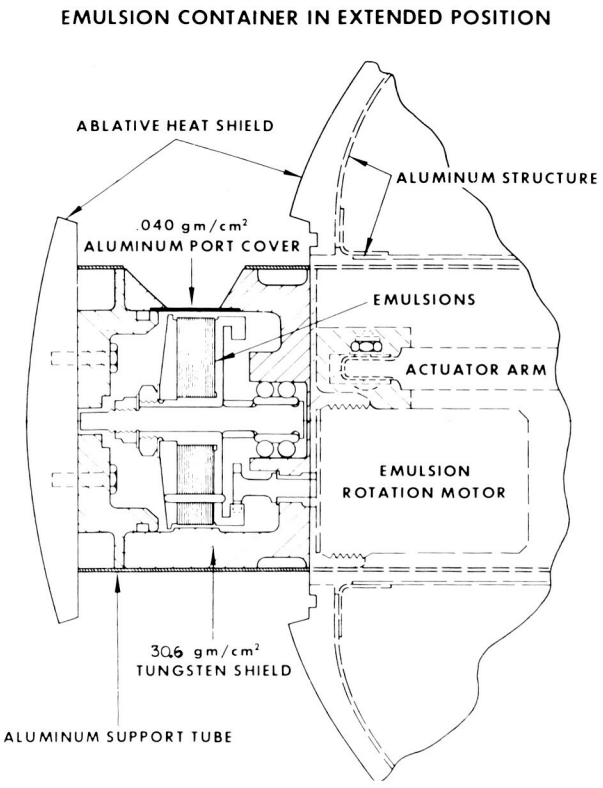


FIGURE 2.—Cutaway side view of the experimental arrangement with emulsion container in extended position.

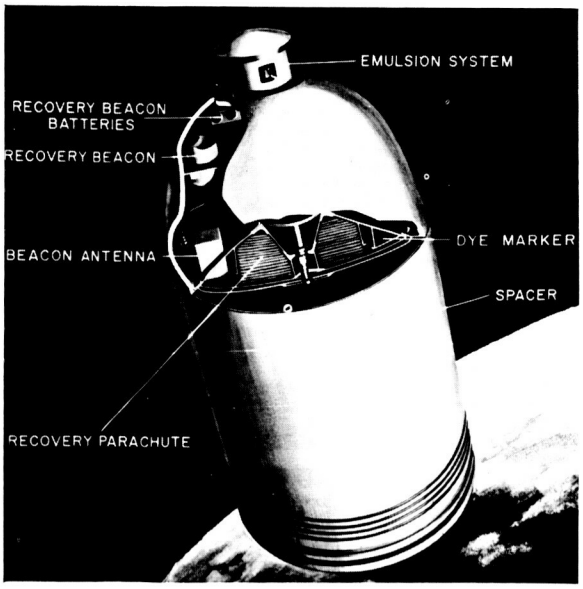


FIGURE 3.—Cutaway view of the nuclear emulsion recovery vehicle with the remainder of the nose cone (emulsion container extended).

old from the 70 Mev in previous experiments down to about 8 Mev in this experiment. Protons of energy ≥ 8 Mev could penetrate the port and be detected in the emulsions, whereas only protons with energy ≥ 145 Mev could penetrate the tungsten and be detected.

The emulsion container was an integral part of the special nuclear emulsion recovery vehicle (NERV). It was designed to permit the 4-inch cylindrical container to be extended through the front of the vehicle so that the emulsion container and its port could be exposed directly to the ambient radiation. Figure 2 is a side view of

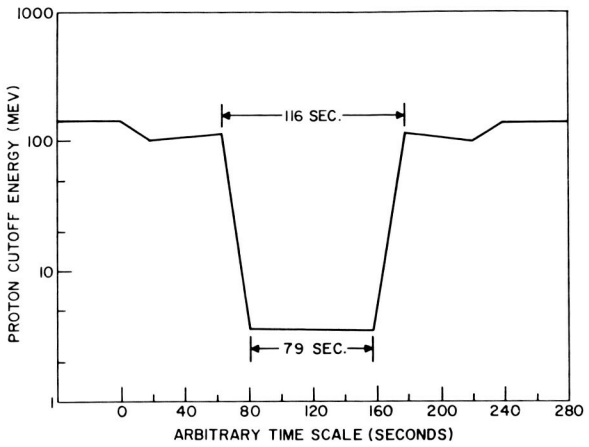


FIGURE 4.—The energy a radial proton must have to penetrate the shield and be detected in the emulsion as a function of time as the emulsion rotates behind the port.

experimental arrangement showing the emulsion container in its extended position. Figure 3 is a cutaway view showing the extended emulsion container and the remainder of the nose cone. The nose cone was covered with an ablation material which, together with the mass of the tungsten, kept the emulsions at $70 \pm 10°$F throughout the flight, including re-entry.

Figure 4 is a plot of the energy a radial proton must have to penetrate the tungsten shield and be detected in the emulsion as a function of time during the passage of a point on the periphery of the emulsion behind the port.

FLIGHT DETAILS

The rocket was launched at 1635 UT, September 19, 1960, from Point Arguello, California.

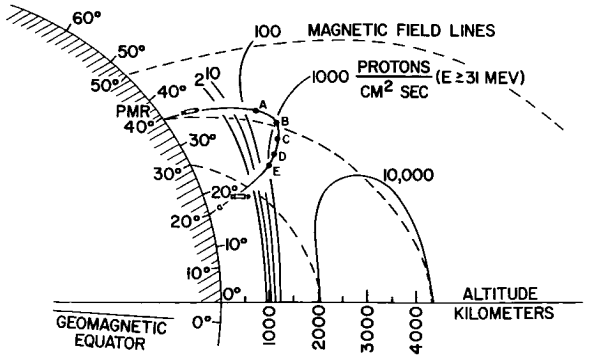


FIGURE 5.—Meridian section containing the trajectory of the NERV rocket. The dashed lines are magnetic lines of force. The solid lines are contours of constant omnidirectional proton intensity based on data from the Iowa group. The lettered points are the locations along the trajectory at which the correspondingly lettered regions of the emulsions in figure 1 were behind the port.

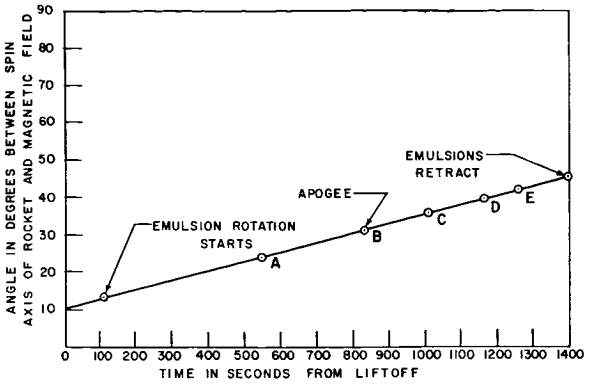


FIGURE 6.—Variation of the angle between the spin axis of the rocket and the local magnetic field vector as a function of time after rocket liftoff.

The launch azimuth of 193° was chosen so that the flight path would parallel a magnetic meridian. The nose cone was recovered at 1923 Ut at 15°N latitude and 126°12'W longitude.

Figure 5 is a meridian section containing the rocket trajectory. The dashed lines are the magnetic lines of force based on the Finch and Leaton [1957] coefficients. The solid lines are the contours of constant integral omnidirectional intensity for protons of energy >30 Mev based on data furnished by the Iowa Group (C. E. McIlwain, private communication). The lettered points are the locations along the trajectory at which the correspondingly lettered regions of the emulsion in figure 1 were behind the port.

Each of the lettered points has been analyzed. Table 1 gives the altitude H, the geomagnetic latitude λ, the value of the scalar magnetic field B, and L (the parameter used by McIlwain [1961] to specify a magnetic shell) at each of these locations along the trajectory.

The payload was spin-stabilized after leaving the atmosphere, the spin axis of the rocket remaining fixed in inertial space. NERV did not contain sensors for the determination of the position or orientation of the rocket, but the payload was tracked by FPS-16 radar to determine the rocket's trajectory. From the analysis of the radar data it was also possible to determine the direction of the spin axis during burning of the fourth stage. With this orientation of the spin axis and the assumption that the spin axis remained fixed in inertial space it was possible to reconstruct the angle γ between the spin axis of the rocket and the local magnetic field vector. The value of γ at each location is given in Table 1. In general, the spin axis was initially nearly parallel to the local magnetic field vector, but

TABLE 1.—*Parameters of the Experiment*

Location	Altitude, km	B, gauss	λ, °N	L, radii earth	γ, deg	$\alpha_1 c'$, deg	αc, deg	$\alpha_{90}°$, deg
A	1600	0.231	33.8	1.79	25	67.2	26.7	29.0
B	1884	0.198	30.5	1.72	32	59.7	28.5	33.7
C	1810	0.196	28.4	1.64	36	61.0	31.8	36.6
D	1600	0.209	26.5	1.54	40	65.2	35.2	39.6
E	1400	0.223	25.3	1.47	42	71.8	38.0	41.3

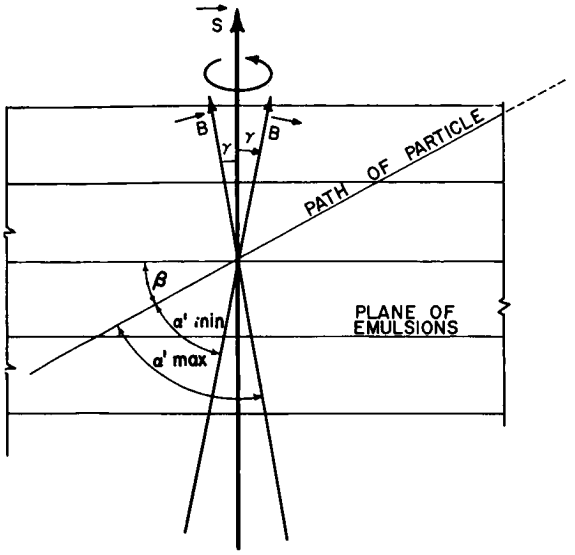


FIGURE 7.—Schematic diagram showing the relation between the dip angle of the track, β; γ, the angle between the spin axis of the rocket **S** and the magnetic field **B**; and α', the pitch angle of the particle.

as the payload traveled along its trajectory this angle steadily increased to 40° at point E as shown in figure 6.

Figure 7 shows the relations between the spin axis of the rocket, the plane of the emulsions, the dip angle of the track relative to the emulsion, and the pitch angle α' of the particles with respect to the local magnetic field vector **B** at the location of the measurement. Viewed in a coordinate system fixed in the spacecraft the magnetic field **B** will appear to rotate about the axis of the rocket. A particle with a given dip angle β in the emulsion will have had a pitch angle α' between $(90°-\beta)\pm\gamma$, depending on the time during a roll period that it entered the emulsion.

At a particular geomagnetic latitude, only a limited range of pitch angles is available, $\alpha_C' \leq \alpha' \leq 90°$, where α_C' is the pitch angle of a particle at the location of the measurement which mirrors at the lower edge of the belt on the particular shell on which the measurement was made. Because the measurements were made at geomagnetic latitudes $\geq 25°$, only a limited range of equatorial pitch angle distribution could be studied in this experiment. The pitch angle at the equator α is related to the pitch angle α' measured at the geomagnetic latitude λ by the relation

$$\alpha = \cos^{-1}\left[1-(1-\cos^2\alpha')\frac{\cos^6\lambda}{(4-3\cos^2\lambda)^{1/2}}\right]^{1/2} \quad (1)$$

α_C and α_{90}, the pitch angles at the equator corresponding to $\alpha'=\alpha_C'$ and to $\alpha'=90°$, respectively, are given in Table 1.

Figure 8 shows how a particular pitch angle distribution at the equator will appear at various geomagnetic latitudes. The exact variation of j with pitch angle and L for these energy ranges has not yet been measured directly. Davis [1962] has found that j is of the form $j=k\sin^3\alpha$ for protons between 1 and 4 Mev. Using the relation derived by Ray [1960], it is found that $j=k\sin^{2.5}\alpha$ by analysis of the omnidirectional intensity J_0. Therefore, figure 8 is a reasonable approximation to the expected pitch angles studied in this experiment.

Figure 8 and Table 1 show that it is possible to explore only a very limited part of the pitch angle distribution at the equator on a flight like this at a high geomagnetic latitude. As Table 1 shows, all the particles observed on this flight have their pitch angles, at the equator, contained in angular intervals that varied from 1.3° to 5°

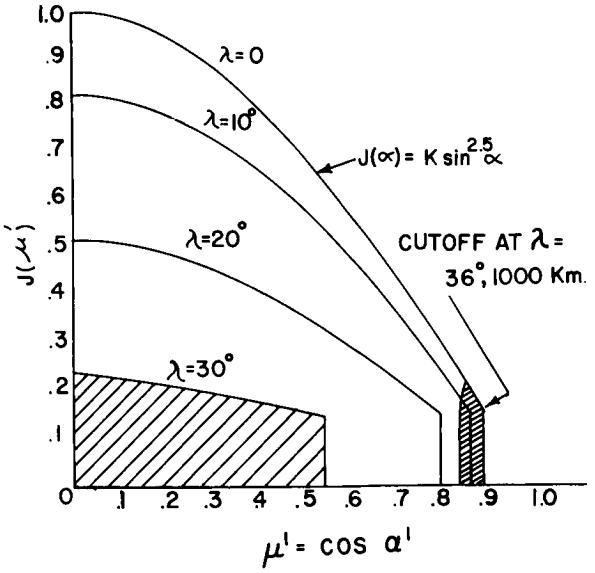


FIGURE 8.—Plot showing how a particular pitch angle distribution changes with geomagnetic latitude. The shaded area under $\lambda=30°$ corresponds to the range of pitch angles accessible in this experiment, and transforms at the equator into the smaller shaded area at $\mu=0.85$.

and centered at pitch angles that varied from 27° to 40° at the geomagnetic equator.

In figure 8, j is plotted as a function of $\mu'=\cos\alpha'$, so that the independent variable μ' runs from 0 to 1. Throughout this paper we will, for simplicity, use μ' as the independent variable rather than α'.

It follows from Liouville's theorem that the unidirectional spectral intensity j remains constant along a trajectory such that $j(\mu)=j(\mu')$, where $\mu=\cos\alpha$ and $\mu'=\cos\alpha'$, and α and α' satisfy equation 1.

For large values of γ or β, there will be a part of the roll period during which the corresponding value of μ' is greater than μ_C'. Figure 9 shows the variation of μ' during a roll period for various values of β and for $\gamma=25°$ and 42°. The horizontal dashed lines labeled 'cutoff' are the values of $\mu_C'=\cos\alpha_C'$ at the particular location. The shaded areas show the period when a given angular interval, in β, was oriented in such a way that the acceptance cone was within the allowed cone of pitch angles. To correct for the effect of the roll on j, the assumption was made that

$$j(\mu)=K \qquad 0\leq\mu'\leq\mu_C'$$
$$j(\mu)=0 \qquad \mu'>\mu_C' \qquad (2)$$

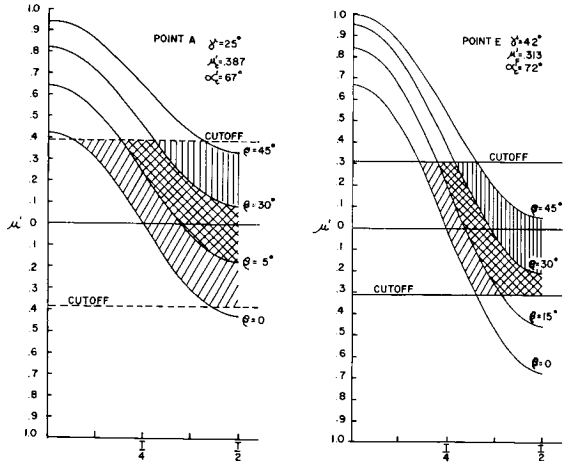


FIGURE 9.—Variation of the pitch angle $\mu'=\cos\alpha'$ for a given value of dip angle β with time during a roll period. The variation is plotted for $\gamma=25°$ and $\gamma=42°$, corresponding to locations A and E, respectively. The cutoff $\mu_{C'}=\cos\alpha_{C'}$ corresponds to an altitude (or B value) on a particular L shell below which satellite data show no particles mirroring.

That is, it was assumed that the pitch angle distribution was isotropic in the allowed cone and zero outside. It is clear from figure 8 that this is a reasonable approximation. The tracks were divided into various dip angle intervals, and the fraction of a roll period f, such that an angular interval was within the allowed cone, was calculated. The flux was then calculated, using an exposure time $t=f\,\Delta t_k$, where Δt_k is the effective time the emulsion was behind the port.

SCANNING CRITERIA

Line scans were made at points A, B, C, D, and E. They were made at 600, 1200, and 2400 microns from the edge of the emulsions; because of blackening it was not possible to scan closer than 600 microns to the edge, but they were made as close to the edge as possible in order to extend the spectrum to the lowest possible energy. However, owing to excess background grains from low-energy electrons, tracks of blob density $b\leq 30$ blobs/100 microns ($\geq$175-Mev protons) could not be detected at 600 microns from the edge. The scanning efficienty for tracks with $b\geq 20$ ($\leq$360-Mev protons) at 1200 microns from the edge as checked by rescans was 100 percent. For a particle at minimum ionization b was about 14. Because of the high background and the low minimum blob density, we could not reliably detect particles at minimum ionization. Therefore, in all the results reported in this analysis, we have used only tracks with blob densities >20.

All tracks that at the scan line were within ±45° of the plane of the emulsion and within ±25° of a radius vector in the plane of the emulsion were recorded and were used to evaluate the integral fluxes. Then tracks that, on the basis of dip measurement at the scan line alone, would have traversed a track length of at least 2.5 mm before leaving the emulsion, independent of energy considerations, were chosen to obtain the energy measurements as described below.

The part of the emulsion labeled *BG*, which was not exposed behind the port, was also scanned with the same criteria to obtain a background correction for the exposed regions of the emulsions.

The scanning to study composition was made in the G-2 emulsion. This allowed tracks to be grain-counted within 200 microns of the end with

good resolution between protons, deuterons, tritons, and α particles down to energies of 14 Mev for deuterons, 16 Mev for tritons, and 40 Mev for α particles.

More than 500 protons were used to establish the grain-density ionization curves for the G-2 emulsions. The grain-density curve for protons was established, and by means of this curve theoretical grain-density versus range curves were constructed for heavier particles by using the value of the grain density of a proton of the same ionization. The theoretical curves were then checked by grain counts on heavy particles from interaction stars. The measured grain densities agreed with the theoretical curve well within the statistical error.

The mass distribution was next determined by selecting, on the basis of dip angle alone, 500 tracks having a possible track length of at least 1 mm. Each track was grain-counted at the scan line and at a fixed distance from the end of the track. Tracks that did not end were discarded, so that a small bias was introduced in favor of heavier particles, owing to the loss of lighter particles by scattering. However, very few partices were eliminated by this criterion, since most of the particles were of low energy and had short residual ranges. In addition, an upper energy cutoff was introduced in the form of a 4-mm residual range cutoff. Therefore, the mass distribution covers protons in the 12- to 31-Mev range, deuterons in the 14- to 42-Mev range, and α particles in the 40- to 125-Mev range.

DETERMINATION OF THE ENERGY

The variation of blob density with energy was determined primarily from scattering measurements and blob counts on ending protons. The assumption was made that the grain density varied with energy according to the Fowler–Perkins law [Fowler and Perkins, 1955]. The blob counts and scattering measurements were made on long, flat tracks. Applying the relation between grain density and blob density $b = ge^{-\delta g}$, where δ is an approximation to the grain size and g in the grain density, a best fit was made to the experimental points taking the minimum blob density as a parameter. A value of $b_0 = 14$ was obtained.

Figure 10 is a plot of energy versus blob density. The crosses are experimental points based on protons that ended in the emulsion; the dots are based on scattering data.

All particles in the scans that had a $b \geq 20/100$ microns and a possible path length in the emulsion at the scan line of at least 2.5 mm, on the basis of dip angle alone, were followed at least 2.5 mm, and a preliminary blob count was made.

Black tracks and gray tracks with $b \geq 40$ were followed until they either ended or interacted with the emulsion. Their energy was determined from their range in emulsion [Atkinson and Willis, 1957]. Approximately 70 percent of the tracks ended in the emulsion at points A, B, and C, and essentially all tracks with energies less than about 70 Mev ended. In contrast, only about 30 percent of the tracks ended at points D and E, emphasizing the relative dirth, at these points, of particles with lower energies.

The energy of particles with $20 \leq b \leq 30$ was

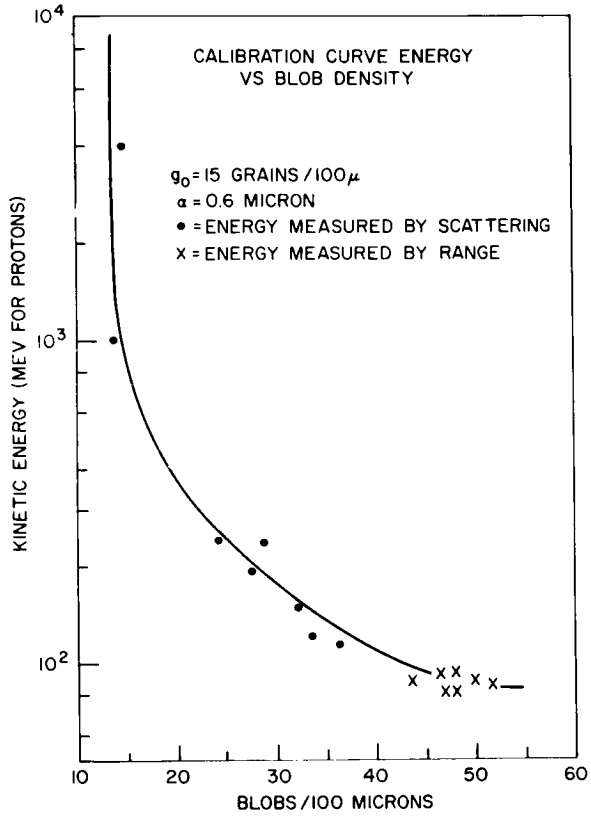


FIGURE 10.—Energy versus normalized blob density.

determined by scattering, and the energy of those with $30 \leq b \leq 40$ by blob density.

It was not possible to obtain a direct cross calibration of all three techniques for energy determination by simultaneously measuring the scattering and blob density as a function of range, since protons with kinetic energies greater than 170 Mev could not be stopped in the emulsion. However, figure 10 shows that the three techniques are self-consistent. If they were not, the experimental points would not lie along the theoretical curve. In the overlapping areas of the three techniques cross checks were made and the energies measured by different techniques were all well within the experimental errors.

CALCULATIONS OF THE FLUX

The flux was determined from the raw data as follows:

Let T be the total time the emulsion container was in the radiation belt ($\sim$1000 sec). Let $j(E)$ be the ambient unidirectional flux in space outside the container, at a position along the trajectory specified by $L_i(t_i)$, $B_i(t_i)$, the MacIlwain coordinates, at a time t_i. The bar over j indicates that it is an average over an energy interval ΔE and a time Δt_i.

Let $\Delta N(E', R, \Theta)\ \Delta E'$ be the number of tracks in the emulsion, in a particular energy interval $\Delta E'$, about the energy E', measured at the scan line whose coordinates in the emulsion are R, Θ, and in the solid angle $\Delta\Omega$, observed in an area ΔA. The track density, $D_{\text{tot}}(E', R, \Theta)$, will be given by

$$
\begin{aligned}
&D_{\text{tot}}(E', R, \Theta)\ \Delta E' \\
&\quad = \frac{\Delta N_{\text{tot}}(E', R, \Theta)\ \Delta E'}{\Delta A\ \Delta\Omega} \\
&\quad = \sum_{i=1}^{k-1} \bar{j}_i(E_s, L_i, B_i)\ \Delta E_s\ \Delta t_i \\
&\qquad + \sum_{\substack{i=k-l \\ i\neq k}}^{k+l} \bar{j}_i(E_i, L_i, B_i)\ \Delta E_i\ \Delta t_i \\
&\qquad + \bar{j}_k,\ (E_k, L_k, B_k)\ \Delta E_k(79) \\
&\qquad + \sum_{i=k+1}^{N} \bar{j}_i(E_s, L_i, B_i)\ \Delta E_s\ \Delta t_i \qquad (3)
\end{aligned}
$$

where the third term is the tracks that came through the port during the 79 seconds the area ΔA was behind the port; the first and fourth terms give the tracks that came in through the 30.6 g/cm^2 of tungsten; and the second term gives the tracks that came in during the time the area ΔA was passing under the edges of the port where E_i is the energy required of a proton in free space to pass through the pertinent thickness of shielding material at time t_i to reach the scan line. E_k is the energy a proton in free space must have in order to pass through the aluminum port cover (0.04 g/cm^2), travel through the emulsion, and have an energy E' at the scan line. E_s is the energy a proton in free space must have had in order to pass through 30.6 g/cm^2 of tungsten and have an energy E' at the scan line. The number of tracks in an area ΔA in the part of the emulsion that did not pass behind the port, the section labeled BG, is given by

$$
\begin{aligned}
&D_{BG}(E', R_{BG}, \Theta_{BG})\ \Delta E' \\
&\quad = D_{BG}(E', R, \Theta)\ \Delta E' \\
&\quad = \frac{\Delta N(E', R_{BG}, \Theta_{BG})\ \Delta E'}{\Delta\Omega\ \Delta A} \\
&\quad = \sum_{i=1}^{N} \bar{j}_i(E_s, L_i, B_i)\ \Delta E_i\ \Delta t_i \qquad (4)
\end{aligned}
$$

Then

$$
\begin{aligned}
[D_{\text{tot}} - D_{BG}]\ \Delta E' &= \bar{j}_k(E_k, L_k, B_k)\ \Delta E_k(79) \\
&\quad + \sum_{\substack{i=k-l \\ i\neq k}}^{k+l} [\bar{j}_i(E_i, L_i, B_i)\ \Delta E_i \\
&\quad - \bar{j}_i(E_s, L_i, B_i)\ \Delta E_s]\ \Delta t_i \\
&\quad - \bar{j}_k(E_s, L_k, B_k)\ \Delta E_s(79) \qquad (5)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\bar{j}_k &= \frac{[\Delta N_{\text{tot}} - \Delta N_{BG}]\ \Delta E'}{79\Delta A\Delta\Omega\ \Delta E_k} \\
&\quad + \bar{j}_k(E_k, L_k, B_k)\ (\Delta E_s/\Delta E_k) \\
&\quad - \frac{1}{79\Delta E_k} \sum_{\substack{k-l \\ i\neq k}}^{k+1} \Delta t_i [\bar{j}_i(E_i, L_i, B_i)\ \Delta E_i \\
&\qquad - \bar{j}_i(E_s, L_i, B_i)\ \Delta E_s] \qquad (6)
\end{aligned}
$$

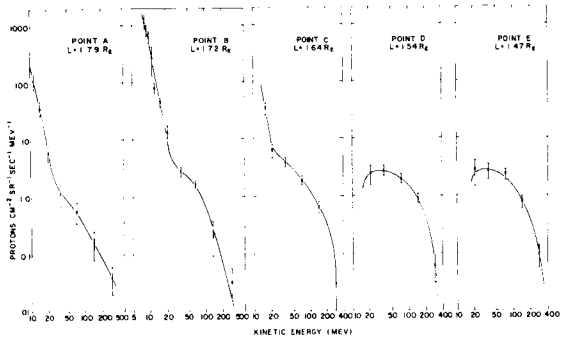


FIGURE 11.—Differential energy spectrums obtained at the five locations.

In practice the first term of equation 6 was calculated at the five points along the trajectory to give the approximate value for $j_i(t_i)$, and then a series of iterations was made to give the final value of j for each of the four points. For $E_k \ll E_s$ the background corrections are small; for $E_k \geq E_s$ the corrections are more important and introduce larger uncertainties into j_k.

RESULTS

The unidirectional-spectral intensity j is plotted in figure 11 separately for each of the five points along the trajectory. These curves represent the absolute value of j for each value of energy and for each position. The data have been corrected for background and for the fraction of the time the acceptance cone was outside the allowed range of pitch angles. Therefore, these curves represent the absolute values of the flux. Figure 12 is a plot of the integral energy spectrum at the five points.

If we take the observed integral unidirectional flux for 30 Mev from figure 12 and integrate the unidirectional flux over solid angle, then $J_0(E \geq 30$ Mev), the integral unidirectional flux of protons of energy ≥ 30 Mev, is obtained and can be compared directly with that observed by Van Allen et al. [McIlwain, 1961] at the same point as shown in Table 2. The agreement between the two is quite good at points B, C, and D and not quite so good at points A and E, although slight errors in the location of the measurement and the rapid variation of the flux with altitude could account for the differences at A and E. For instance, if B at point E were in error by 0.01 gauss, there would be a factor-of-3 variation in the expected integral flux.

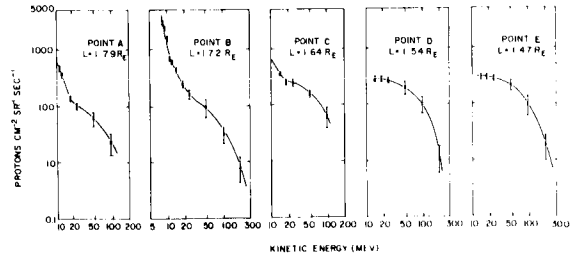


FIGURE 12.—Integral energy spectrums obtained at the five points.

TABLE 2.—*Comparison of Counter and Emulsion Fluxes*

Location	$J(\geq 31$ Mev), cm^2 sr sec	$J_0(31$ Mev), cm^2 sec	J_0*(31 Mev), cm^2 sec
A	90	440	220
B	180	1160	1000
C	220	1360	1600
D	230	1230	1290
E	250	1000	390

*Satellite data from C. E. McIlwain, private communication.

The lack of particles below 24 Mev at points D and E is real, since energies down to 12 Mev were detectable within the scan criteria. However, the numbers of particles found were so few that the intensity fell below the scale of the graphs.

DISCUSSION OF THE RESULTS

The most pronounced feature of the various spectrums is the systematic variation in the shape of the spectrum with position. This unexpected result indicates that, in the energy region below 30 Mev, the galactic neutron albedo theory cannot account for the shape of the spectrum or its behavior with position. This theory [Freden and White, 1962] would predict a dip in the energy spectrum below 30 Mev due to an increased neutron absorption at these energies by the nitrogen and oxygen in the upper atmosphere. This absorption mechanism would account for

the maximum in the energy spectrum at 30 Mev at points D and E and the plateau in the spectrum at the same energy that is observed at point B. However, although this theory can explain the paucity of protons below 30 Mev at the low L values (points D and E), it cannot explain the copious numbers of protons below 30 Mev observed at the higher L values (points A, B, C) in this and other experiments [Bame et al., 1961; Davis, 1962].

There are several mechanisms by which such a steep spectrum can be obtained at these energies; however, the high intensities of particles require a stronger source of protons than could be provided by the galactic cosmic ray neutron albedo spectrum measured by Hess et al. [1959]. There might be a serious dilemma in deciding among the various possible explanations of the combined steep spectrum and high intensities in the 8- to 30-Mev region were it not for the fortunate circumstance that the trajectory of the rocket passed through the quasi-boundary between points C and D where this component of the spectrum disappeared.

This feature gives an additional criterion with which to seek the source of protons giving rise to this form of spectrum. From figure 12 it can be seen that the slope of the proton spectrum between 16 and 48 Mev changes rapidly beginning at an L value of about 1.55 with relatively few of the low-energy protons present below this value of L and a great abundance of them above it. Thus we must look for an intense source of low-energy protons that has access only to higher magnetic shells. This type of spectrum could reasonably be expected from a source of protons that come from the decay of neutrons arising from the interaction of low-energy solar protons with the earth's atmosphere over the polar regions. A sufficiently intense source of such particles is known to arrive at the earth after certain types of solar flares [McDonald, 1962].

Such an explanation of the spectral variations at low energies was suggested in the preliminary results of this experiment [Naugle and Kniffen, 1961; Lenchek and Singer, 1962].

Subsequently, quantitative calculations to determine the feasibility of such an explanation were made by Lenchek [1962]. He calculates from a time-averaged solar cosmic ray spectrum the neutron spectrum to be expected as the solar corpuscular particles collide with the earth's upper atmosphere. Assuming adiabatic trapping, he calculates the injection coefficients for the various locations, the pitch angles, and the lifetimes, and he shows that this spectrum could be an essentially time-invariant feature of the trapped proton spectrum in the allowed regions. His calculated spectrum and its variation with position are in fair agreement with the spectrum observed at points A, B, and C in this experiment, providing good evidence in support of this type of source for the steep low-energy component since it is the only apparent explanation of the absence of low-energy protons at the lower L values. Conclusive evidence for this source, however, cannot be obtained until there is a complete study of the variation of the trapped proton energy spectrum as a function of the pitch angle at the equator, or any of the alternative measurements that can provide these data.

Bame et al. [1961] and Davis [1962] have shown that, at high L values, the proton energy spectrum continues to rise steeply down to about 0.1 Mev. Freeman [1962] has observed an extremely high flux of very-low-energy protons in regions of L value from which solar protons are excluded. The intensity of this low-energy component is so high that it appears impossible, with reasonable lifetimes, to account for it from either the galactic or solar cosmic ray source. Thus, it appears that there may be at least three sources for trapped protons: galactic cosmic rays for the high-energy protons at all L values; solar cosmic rays for the low-energy (1- to 30-Mev) protons at high L values; and some as yet unknown accelerating mechanism for the very-low-energy (0.1- to 1-Mev) protons.

Because there is such a large background correction to the data above about 100 Mev the errors on the data points are quite large. This is due to the collection over the entire periphery of the emulsion of particles that penetrate the tungsten shielding during the entire time of flight. Thus it is only possible to say that at higher energies the gross shape of the spectrum remains essentially the same and there is evidence for a decrease of intensity at the higher L value. How-

ever, there are not sufficient data to give new information about the loss mechanism for trapped particles.

At comparable positions the data at the high energies are in agreement within a factor of 2 with previous results [Freden and White, 1962; Armstrong and Heckman, 1962], which is as good as should be expected in view of the differences in trajectory and exposure technique.

In a study of 500 tracks to investigate particle composition in the proton range interval corresponding to 12 to 31 Mev, no particles of mass greater than the proton were found. For a sample of this size there is only a 5 percent probability that the observer would expect three heavier particles and obtain none. Thus the heavier particles due both to primaries and to secondaries from interaction in the tangsten make up less than 0.6 percent of the entire population of observed tracks, and so this can safely be taken as an upper limit to the number expected from primaries alone.

SUMMARY

1. At L values above about $1.6R_E$ the slope of the proton energy spectrum below about 30 Mev is very steep compared with predictions of the galactic cosmic ray neutron albedo theory and is most probably due to solar cosmic ray neutron albedo; however, the flux of trapped protons with energies greater than about 30 Mev is in agreement with the prediction of this theory.

2. The shape and flux of the spectrum vary with latitude, points A and D being at the same altitude.

3. At comparable positions in the radiation region the results are in reasonable agreement with previous measurements.

4. On the basis of these and subsequent data, it appears that there must be at least three sources of trapped protons.

5. The flux of geomagnetically trapped particles heavier than protons in the proton energy range of 12 to 30 Mev is less than 0.6 percent of the total flux, on the date and at the locations studied in this experiment.

ACKNOWLEDGMENTS

An experiment like this, which required recovery at sea, would not have been successful without the superior performance and cooperation of so many people that, because of the limitations of space, we cannot acknowledge them individually. Charles Campbell, project engineer, and Gerald Longanecker, assistant project engineer, each made major contributions which ensured the success of the project.

REFERENCES

Armstrong, A. H., Harrison, F. B., Heckman, H. H., and Rosen, L., Charged particles in the inner Van Allen radiation belt, *J. Geophys. Res.*, **66**, 351–357, 1961.

———, and Heckman, H. H., Energy spectrum of geomagnetically trapped protons, *J. Geophys. Res.*, **67**, 1255–1262, 1962.

Atkinson, J. H., and Willis, B. H., High-energy particle data, vol. 2, *UCRS–2426 (rev.)*, June 1957.

Bame, S. J., Conner, J. P., and Hill, H. H., Protons in the outer Van Allen Belt, First Western National Meeting of the American Geophysical Union, 1961.

Davis, L. R., Low-energy trapped protons, Forty-Third Annual Meeting of the American Geophysical Union, 1962.

Dessler, A. J., Effect of magnetic anomaly on particle radiation trapped in the geomagnetic field, *J. Geophys. Res.*, **64**, 2239–2252, 1959.

Dragt, A. J., Effect of hydromagnetic waves on the lifetime of Van Allen radiation protons, *J. Geophys. Res.*, **66**, 1641–1649, 1961.

Finch, H. F., and Leaton, B. R., *Monthly Notices Roy. Astron. Soc. Geophys. Suppl.*, **7**(6): 314–317, November 1957.

Fowler, P. H., and Perkins, D. H., Measurement of ionization in nuclear emulsions, *Phil. Mag.*, **46**, 587–592, 1955.

Freden, S. C., and White, R. S., Particle fluxes in the inner radiation belt, *J. Geophys. Res.*, **65**, 1377–1383, 1960.

———, and ———, Trapped proton and cosmic ray albedo neutron fluxes, *J. Geophys. Res.*, **67**, 25–29, 1962.

Freeman, J. N., Detection of an intense flux of low-energy protons or ions trapped in the inner radiation zone, *J. Geophys. Res.*, **67**, 921–928, 1962.

Gold, T., Motions in the magnetosphere of the earth, *J. Geophys. Res.*, **64**, 1219–1224, 1959.

Hess, W. N., Van Allen belt protons from cosmic-ray neutron leakage, *Phys. Rev. Letters*, **3**, 11–13, 1959.

HESS, W. N., PATTERSON, H. W., and WALLACE, R., Cosmic-ray neutron spectrum, *Phys. Rev.*, **116,** 445–457, 1959.

IMHOF, W. L., SMITH, R. V., and FISHER, P. C., Proton measurements in the Van Allen belt, 1962 Spring Meeting of the American Physical Society, Washington, D.C.

LENCHEK, A. M., On the anomalous component of low-energy geomagnetically trapped protons, *J. Geophys. Res.*, **67,** 2145–2157, 1962.

———, and SINGER, S. F., Geomagnetically trapped protons from cosmic ray albedo neutrons, *J. Geophys. Res.*, **67,** 1263–1287, 1962.

MCDONALD, F. B., editor, *Goddard Space Flight Center Proton Manual,* contributed by C. E. Fichtel, D. E. Guss, H. H. Malitson, K. G. McCracken, K. W. Ogilvie, and W. R. Webber (see also references therein), 1962.

MCILWAIN, C. E., Coordinates for mapping the distribution of magnetically trapped particles, *J. Geophys. Res.*, **66,** 3681–3691, 1961.

NAUGLE, J. E., and KNIFFEN, D. A., The flux and energy spectra of the protons in the inner Van Allen belt, *Phys. Rev. Letters,* **7,** 3–6, 1961.

PIZELLA, G., MCILWAIN, C. E., and VAN ALLEN, J. A., Time variations of intensity in the earth's inner radiation zone, October 1959 through December 1960, *J. Geophys. Res.*, **67,** 1235–1253, 1962.

RAY, E. C., On the theory of protons trapped in the earth's magnetic field, *J. Geophys. Res.*, **65,** 1125–1134, 1960.

SINGER, S. F., 'Radiation belt' and trapped cosmic-ray albedo, *Phys. Rev. Letters,* **1,** 171–173, 1958.

———, On the cause of the minimum in the earth's radiation belt, *Phys. Rev. Letters.* **3,** 188–190, 1959.

VAN ALLEN, J. A., LUDWIG, G. H., RAY, E. C., and MCILWAIN, C. E., Observations of high-intensity radiation by satellites 1958α and γ, *Jet Propulsion,* **28,** 588–592, 1958.

VERNOV, S. N., GRIGOROV, N. L., IVANENKO, I. P., LEBEDINSKII, A. I., MURZIN, V. S., and CHUDAKOV, A. E., Possible mechanism of production of 'terrestrial corpuscular radiation' by the action of cosmic rays, *Dokl. Akad. Nauk SSSR,* **124,** 1022–1025, 1959.

WELCH, J. A., and WHITAKER, W. A., Theory of geomagnetically trapped electrons from an artificial source, *J. Geophys. Res.*, **64,** 909–922, 1959.

WENTZEL, D. G., Hydromagnetic waves and the trapped radiation, *J. Geophys. Res.*, **66,** 359–369, 1961.

SOLAR PROTON SPECTRUMS IN THE EVENTS OF NOVEMBER 12 AND 15, 1960

K. W. OGILVIE AND D. A. BRYANT

Goddard Space Flight Center

Freier and Webber have demonstrated that an exponential rigidity spectrum provides a good empirical fit to solar proton data over a wide range of rigidity and during widely different events. In this paper we discuss the conclusions they have drawn about the events of November 12 and 15, 1960, during which a series of rocket flights was made by the Goddard Space Flight Center. We find that for the November 12 event an exponential rigidity spectrum does not fit the observations, owing to the presence of a relatively large number of low-energy protons, and that data obtained late in the November 15 event are consistent with an exponential rigidity spectrum. We suggest that the difference in character of the spectrums observed in these two events is due to a flux of low-energy protons arriving at the earth on November 12 with the solar plasma responsible for concurrent geomagnetic disturbances. The flights made late in the November 15 event were during a geomagnetically quiet time.

INTRODUCTION

Freier and Webber [1963] have proposed an exponential rigidity form for the spectrums of solar particles. In deriving this form they used the results of balloon flights, neutron monitor and riometer observations covering an energy range from about 15 Mev to several bev.

The suggestion that high-energy particles from the sun have a rigidity spectrum of the form $J=J_0 \exp(-P/P_0)$, where J is the intensity above rigidity P, J_0 is the intensity above zero rigidity, and P_0 is the characteristic rigidity, which is a function of time during a given event, has led us to re-examine (and rework) the results we obtained during the solar proton events of November 12 and 15, 1960. The purpose of this work is to test the fit of an exponential spectrum at low energies by means of direct particle observations.

In our original analysis (Ogilvie et al., 1962; Davies and Ogilvie, 1962] we assumed the spectrum to be a power law in kinetic energy; hence we are interested in determining whether the exponential form provides a better fit to our data, and, if so, the extent to which the intensity values are changed.

ANALYSIS

During the November 1960 solar proton events several rockets were fired from Fort Churchill by the Goddard Space Flight Center. These rockets carried both particle counters and nuclear emulsions (characteristics of the particle counters are listed in Table 1).

TABLE 1

Detector	Lower energy limit	Upper energy limit
Geiger counter	30 Mev	—
CsI Sc. counter	2 Mev	160 Mev
ZnS Sc. counter	0.2 Mev	4.5 Mev

We shall be concerned mainly with the observations obtained from the 0.25-g/cm^2 CsI scintillator. An integral pulse height analysis was performed on the pulses from this counter by altering the collector impedance of the phototube using a motor-driven switch followed by an amplifier with a fixed trigger level. This system has been fully

described before, but it is necessary to re-emphasize that such a system introduces overlapping energy intervals with none of the limits common to all levels. An appropriate method of analysis is thus to assume a spectral form and then to compute the ratios of the rates in each interval as a function of one or two parameters and to compare the computed and observed ratios. The energy levels are illustrated in figure 1. The solid curve shows energy loss as a function of energy for normally incident particles. All angles of incidence are taken into account in the analysis. Since we have four levels, three ratios are defined, and to get a meaningful we can use only two parameters. In our earlier paper we used n, the exponent in an integral power-law spectrum $N=N_0E^{-n}$, and a low-energy cutoff Ec. The low-energy cutoff was introduced to account for the observation that the counting rates on the two lowest energy levels were often almost equal, showing that there were few particles in the energy region between their lower limits. We obtained in this way satisfactory fits for the three rocket shots into the November 12 event, and somewhat less satisfactory fits for the four rocket shots fired into the November 15 event. This paper discusses the results we have obtained by recalculating the ratios of the counting rates of the various energy levels assuming both a powerlaw spectrum and an exponential rigidity spectrum. For the exponential rigidity analysis we have used P_0, the characteristic rigidity parameter discussed by Freier and Webber, and W_p, the proportion of α particles to protons in a given rigidity interval. For the power-law tests we have used separately two sets of two parameters: the exponent n and a low-energy cutoff Ec, and the exponent n and the proportion W_E of α particles to protons in a given energy per nucleon interval.

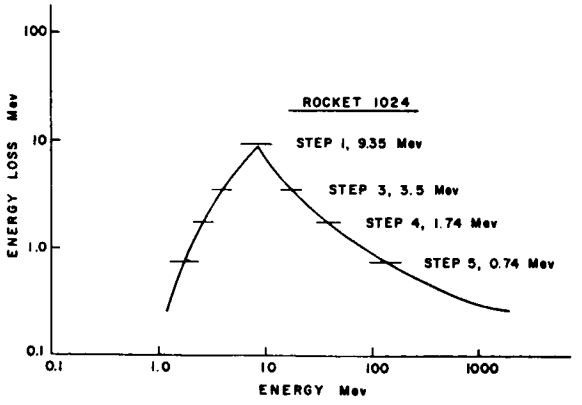


FIGURE 1.—Energy loss in 0.25-g/cm² CsI as a function of proton energy for normally incident particles. The integral energy loss discrimination levels for each step are shown by horizontal lines.

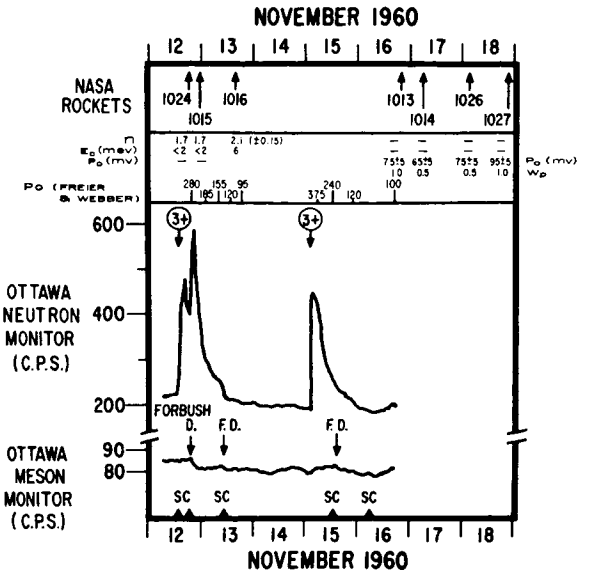


FIGURE 2.—The events of November 12 and 15, 1960, showing the times of NASA rocket flights, results obtained by Frier and Webber and the present work, and solar and terrestrial events.

RESULTS

Figure 2 illustrates several important features of the events of November 12 and 15. It shows the Ottawa neutron monitor and meson monitor counting rates and the occurrence of solar flares and magnetic storm sudden commencements. Also shown are the times of firing of the NASA rockets and the values of the parameters which best fit the counter data, together with values of P_0 deduced by Freier and Webber.

November 12 Event

All the observations on November 12 took place during the active part of this complicated event. A class 3 flare at 1000 UT on November 10, which had no type IV emission and gave rise to no detect-

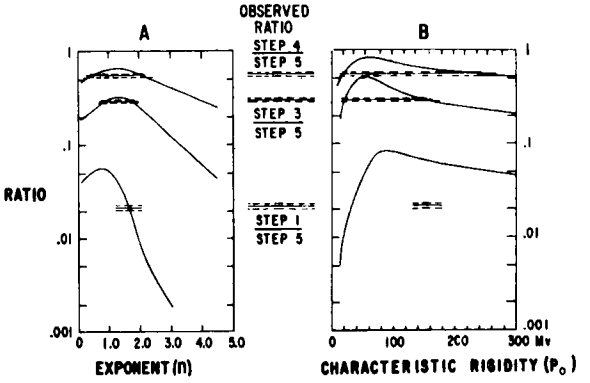


FIGURE 3.—The analysis of the first rocket flight (NASA 1024, 1840 U.T., November 12, 1960). The horizontal lines represent the observed ratios, with errors, and curves show computed values of the ratios as a function of exponent n (diagram A) and characteristic rigidity P_0 (diagram B).

able particles, produced the sudden commencement at 1844 UT on November 12. By coincidence, this sudden commencement occurred during the arrival at the earth of solar protons emitted by the flare of 1320 UT on November 12.

The first rocket (NASA 1024) was fired at 1840 UT on November 12, four minutes before the sudden commencement. The second (NASA 1015) was fired approximately five hours later at 2332 UT, about $2\frac{1}{2}$ hours after the second of two large increases in the earth's field. The third rocket (NASA 1016) was fired at 1603 UT on November 13 during the recovery from the 1500-γ negative excursion that occurred at 1023 UT on November 13. The disturbed magnetic conditions at the time of these firings show that the magnetosphere was at these times immersed in a solar plasma stream.

Good fits to the data from the first two rockets were obtained using integral power-law spectrums with exponent $n=1.7\pm0.1$. The analysis of the first of these flights (NASA 1024) is illustrated in figure 3. We show the computed ratios of the rates on the three steps, 1, 3, and 4 to the rate on step 5 for two spectral forms. In diagram A the assumed spectrum is a power law in kinetic energy and in diagram B an exponential law in rigidity. The highest solid line in each diagram is the ratio of the rate on step 4 to the rate on step 5, the middle solid line is the ration of the rate on step 3 to that on step 5, and the lowest line is the ratio of the rate on step 1 to that on step 5. In general the observed ratios, which are shown as horizontal lines (with errors), cut the corresponding computed curve twice. In diagram A an intersection of the observed ratios with the corresponding computed curves occurs for $n=1.7\pm0.1$, indicating a fit to a power law with this integral exponent. In diagram B a single value of characteristic rigidity consistent with the observed and computed ratios cannot be found. There is a set of intersections between 20 Mv and 40 Mv; we reject this poor fit, since it is inconsistent with other measurements. The results from the third rocket can be fitted with a power-law spectrum of exponent $n=2.1\pm0.15$ and $Ec=6.0$ Mev, and somewhat less satisfactorily by an exponential rigidity spectrum with $P_0=65$ Mv. These results are essentially the same as those given in our first paper [Ogilvie et al., 1962].

We have re-examined the evidence for the low-energy protons (from 0.2 to 4.5 Mev) detected by the ZnS scintillation counter and find that the spectrum given in Figure 9 of our earlier paper [Ogilvie et al., 1962] remains unchanged.

The November 15 Event

The times of rocket firings into this event are shown in figure 2. During the November 15 event all firings were made during magnetically quiet times, between 41 and 91 hours after the flare. An exponential rigidity spectrum provides a good fit to the data from these flights. The analysis of the first of these flights (1951 UT, November 16) is presented in figure 4 as an example. The arrangement of figure 4 is similar to that of figure 3. The

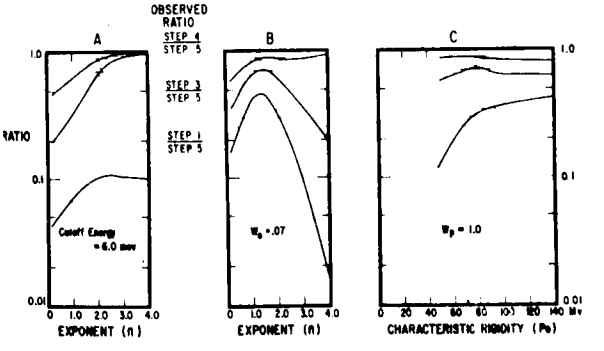


FIGURE 4.—The analysis of the first rocket flight into the November 15 event (NASA 1013, 1951 U.T., November 16, 1960). (For explanation see text.)

fit to the exponential rigidity spectrum is shown in diagram C. We find that, in order to make the fit, it is necessary to postulate a value of W_p close to unity. This high proportion of α particles is in agreement with measurements made at higher energies earlier in the same event [Freier, 1963]. Diagram A shows that a power-law spectrum with a low-energy cutoff does not fit the data even when the cutoff energy E_c is as high as 6 Mev. Diagram B of figure 4 shows that a power-law spectrum with a value of W_E of 0.07 provides a fit as close as that provided by the exponential rigidity spectrum. This value of W_E is consistent with measurements made on the same rocket by Biswas and Fichtel (private communication).

In the three later flights into the November 15 event the fit given by the exponential rigidity spectrum is decisively better than that given by the power law for any proportion of α particles. As can be seen from the data given in Figure 2, direct comparison of the value of P_0 obtained from this experiment with the value of P_0 obtained by Freier and Webber can be made only in one case. Although the disparity in this case is outside that expected from statistical fluctuations, we note that the values of P_0 we observed on November 17 and 18 are in agreement with the values to be expected from the lowering of P_0 with time on November 15 and 16 as observed by Freier and Webber.

We now compare in Table 2 the intensity above 10 Mev deduced by means of the exponential rigidity spectrum with that previously obtained by means of a power law in kinetic energy. This

TABLE 2.—*Comparison of Intensity Values in the November 15 Event Obtained by Two Different Analyses*

Rocket	Time	Intensity: p/cm² sec ster>10 Mev	
		Power-law analysis	Rigidity analysis
NASA 1013	1951, Nov. 16	530±50	430±70
NASA 1014	0600, Nov. 17	200±40	460±100
NASA 1026	0339, Nov. 18	150±20	140±30
NASA 1027	2139, Nov. 18	45±5	25±5

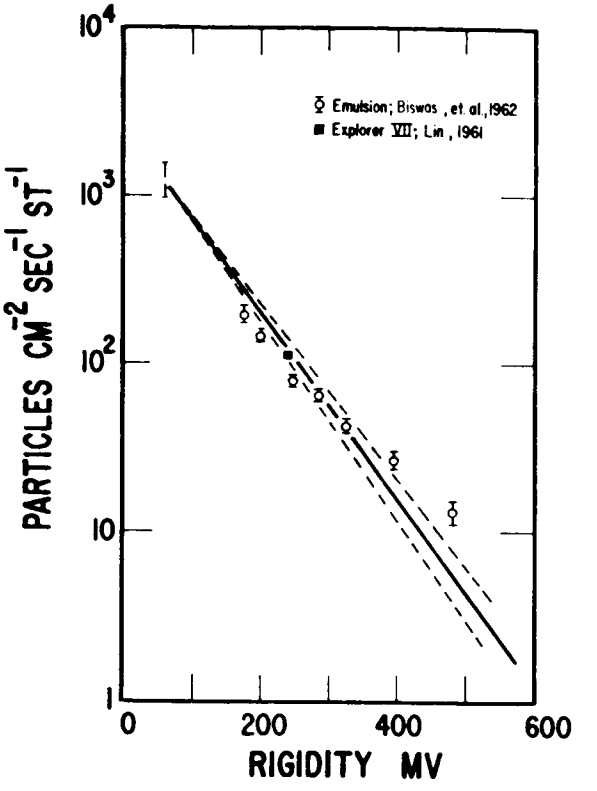


FIGURE 5.—A comparison of the emulsion and counter results for the spectrum on flight NASA 1013, and an intensity derived from counter measurements on Explorer 7.

comparsion is made at 10 Mev, since particles of this energy can produce counts on all steps(taking into account particles whose trajectories are inclined to the axis of the crystal).

Table 2 shows that the rigidity analysis which gives the better fit to the data yields intensity values within a factor of 2 of those given by the power-law analysis. As figure 5 shows, the rigidity analysis provides a good fit to data obtained at other energies.

In interpreting the counting rates of the ZnS counter in the November 15 event, we now think it possible that a large proportion of these counts could have been caused by higher-energy particles passing through the cathode of the phototube, and this therefore throws some doubt on the rising parts of the spectrums below 1 Mev indicated in an earlier paper [Davis and Ogilvie, 1962]. It should be mentioned here that this is not the case for the November 12 event, for which direct

evidence from the absorption curves leaving and entering the atmosphere confirms the existence of low-energy particles.

CONCLUSIONS

1. Neither the exponential rigidity spectrum nor any other functional form we have tried will at all times represent the low-rigidity end of the solar proton spectrum.

The fact that the exponential rigidity form does not always fit does not conflict with any of the observations used by Freier and Webber since the detectors from which they obtain low-energy particle information were riometers for which the differential sensitivity to protons falls rapidly with decreasing energy below about 15 Mev (D. E. Guss, private communication). A consequence of this result is that an extrapolation of the spectrum measured above 15 Mev down to zero rigidity provides a value of J_0 which, although it is a useful parameter to describe the high-energy spectrum, is not necessarily the same as the intensity above zero rigidity.

2. We suggest that the low-energy protons observed on November 12 were associated with the arrival at the earth of the solar plasma responsible for the very disturbed geomagnetic conditions at that time.

ACKNOWLEDGMENTS

We acknowledge the contribution of L. R. Davis, who originated the experiment and designed the detectors. We are also grateful to Drs. C. E. Fichtel and F. B. McDonald for helpful discussions.

REFERENCES

Biswas, S., Fichtel, C. E., Guss, D. E., and Waddington, C. J., Hydrogen, helium, and heavy nuclei from the solar event on November 15, 1960, *GSFC Reprint X-611-62-235*, 1962.

Davis, L. R., and Ogilvie, K. W., Rocket observations of solar protons during the November 1960 events, 2, *J. Geophys. Res.*, **67**(5): 1171, 1962.

Freier, P. S., Emulsion measurements of solar alpha particles and protons, *J. Geophys. Res.*, **68**(7): 1805, 1963.

———, and Webber, W. R., Exponential rigidity spectrums for solar flare cosmic rays, *J. Geophys. Res.*, **68**(6): 1605, 1963.

Lin, W. C., Observations of galactic and solar cosmic rays from 13 October 1959 to 17 February 1961 with Explorer 7 satellite, *State Univ. Iowa Rept. SUI-61-16*, 1961.

Ogilvie, K. W., Bryant, D. A., and Davis, L. R., Rocket observations of solar protons during the November 1960 events, 1, *J. Geophys. Res.*, **67**(2): 1929, 1962.

ON THE MOTION OF CHARGED PARTICLES IN THE GEOMAGNETIC FIELD

ERNEST C. RAY*

Goddard Space Flight Center

We construct an approximate first integral of the equations of motion of a charged particle in a static magnetic field. The integral is analogous to the Størmer integral, but it is not required that the field have axial symmetry in order for the integral to be exact. The integral reduces to that of Størmer when the field does possess axial symmetry. It is used to obtain results concerning trapped particles and cosmic rays previously derived by various people using a variety of techniques. The new technique can readily be extended to cover situations not previously tractable.

INTRODUCTION

H. Grad[1] has pointed out the utility of the representation of a magnetic field by two scalar functions α and β where the vector potential is given by

$$\mathbf{A}=\alpha\nabla\beta. \tag{1}$$

This vector potential satisfies the gauge condition $\mathbf{A}\cdot\mathbf{B}=0$ rather than the usual $\nabla\cdot\mathbf{A}=0$. Whenever any currents present flow at right angles to $\mathbf{B}$, we can also represent $\mathbf{B}$ with two other functions μ and V where

$$\mathbf{B}=\mu\nabla V. \tag{2}$$

We shall use the three functions α, β, and V as a new set of coordinates with which to express the equations of motion of a charged particle in a static magnetic field, and assume that β is a cyclic coordinate. This permits the construction of a first integral analogous to that of Størmer. In practice, β is apparently not cyclic, although in one case of considerable geophysical interest it is so approximately and with good accuracy. All results obtained from this integral are consequently also approximate, and this word will not be tediously repeated for each.

*National Academy of Sciences fellow on leave from the department of Physics and Astronomy, the State University of Iowa, Iowa City, Iowa. Present address: Center for Radiophysics and Space Research, Cornell University, Ithaca, New York.

For trapped particles, α at the guiding center is shown to be a constant of the motion. From its definition, it is constant along a line of force. In a dipole field it becomes $\cos^2\lambda/r$ where λ is geomagnetic latitude and r is radial distance. These three facts result in its identification with L^{-1}, where L is McIlwain's parameter.[2]

For cosmic rays, the integral leads to a theory of cutoffs in the Størmer sense. It is shown that previously obtained results in this direction at high latitudes follow from the new theory.[3] In addition, it is shown that cutoffs along any particular line of force, in a particular cone of directions of half-angle $\pi/2$, are independent of position along the line. This fact has ben previously verified numerically for a selected set of lines of force for particles traveling tangent to the line of force at arrival.[3] Webber's impact zone theory[4] is an evident consequence of this new first integral.

All of these results give reasonable confidence in the good accuracy of the integral, in the case where the magnetic field is the usual one obtained by harmonic analyses of surface data, with or without external sources which possess axial symmetry. Under these circumstances one can then use the integral to calculate Størmer cutoffs at all latitudes and to considerably simplify the study of impact zones for latitudes too low for Webber's method to apply. In addition, it is reasonable to suppose that a main cone and

penumbra analogous to those in a dipole field exist.

An unexplored question of considerable interest is the extent to which the accuracy is preserved when the magnetic field at great distances departs strongly from axial symmetry. This question is important for the case where the earth is in the path of a solar wind with the field at large distances being swept back in the direction away from the sun.

THE MAGNETIC FIELD

We review the description of the magnetic field.[1]

Suppose that one adds the gradient of a scalar function to the vector potential such that the new vector potential satisfies the condition $\mathbf{A}\cdot\mathbf{B}=0$, i.e.,

$$\mathbf{A}\cdot\nabla\times\mathbf{A}=0. \tag{3}$$

This is the well known condition that the differential form $\mathbf{A}\cdot\mathbf{dl}$ possess an integrating factor, α. Then it follows that

$$\mathbf{A}=\alpha\nabla\beta. \tag{4}$$

If the magnetic field possesses axial symmetry, one can put

$$\alpha=r\sin\vartheta A_\varphi(r,\vartheta)$$
$$\beta=\varphi, \tag{5}$$

where r, ϑ, φ are spherical coordinates with the symmetry axis being the z-axis of the coordinate system, and A_φ the φ component of the vector potential, the only component which does not vanish. In the dipole case, (5) becomes

$$\alpha=M\sin^2\vartheta/r$$
$$\beta=\varphi \tag{6}$$

where M is the dipole moment.

In the general case, we have

$$\mathbf{B}=\nabla\alpha\times\nabla\beta. \tag{7}$$

From this it is obvious that both α and β are solutions of the partial differential equation for f,

$$\mathbf{B}\cdot\nabla f=0. \tag{8}$$

Since the characteristic equations of (8) are

$$\mathbf{dl}=(dl/B)\mathbf{B}, \tag{9}$$

it is evident that

$$\alpha=\text{const.}$$
$$\beta=\text{const.} \tag{10}$$

are the equations of a line of force.

There is another representation of a magnetic field which, in many situations, is available. Assume that throughout any region where it is desired to compute the magnetic field, any currents present flow at right angles to the magnetic field, so that if $\mathbf{j}$ is the current density,

$$\mathbf{j}\cdot\mathbf{B}=0 \tag{11}$$

Except for brief transients this is normally true in plasmas. Then, by one of Maxwell's equations, and assuming that the displacement current vanishes, (11) becomes

$$\mathbf{B}\cdot\nabla\times\mathbf{B}=0. \tag{12}$$

Consequently there exist two scalar functions μ and V such that

$$\mathbf{B}=\mu\nabla V. \tag{13}$$

It is obvious that

$$\nabla V\cdot\nabla\alpha=\nabla V\cdot\nabla\beta=0, \tag{14}$$

and while it appears that in general $\nabla\alpha$ and $\nabla\beta$ are not perpendicular to each other, α and β are nevertheless independent functions, so that it is possible to use α, β, V as a set of coordinates. This possibility will be exploited in the next section. In the event that one needs a name for the vector ∇V, it is suggested that one put

$$\mathbf{H}=\nabla V. \tag{15}$$

When $\mathbf{j}=0$ in a region, we of course have $\mu=1$ there.

The following method of computing $\mathbf{A}$ when it satisfies (3) is of interest.

Let $\mathbf{A}'$ denote the vector potential to such a gauge that $\nabla\cdot\mathbf{A}'=0$, and presume that $\mathbf{A}'$ is known as a function of position. Choose a particular surface $V=$ constant as a reference surface and denote any quantity evaluated there with a subscript 0, as $\mathbf{A}_0$ for example. Let $\mathbf{A}$ be the vector potential to such a gauge that (3) is satisfied, and define a function χ so that

$$\mathbf{A}=\mathbf{A}'-\nabla\chi. \tag{16}$$

Then (3) becomes

$$\mathbf{B}\cdot\mathbf{A}'=\mathbf{B}\cdot\nabla\chi. \tag{17}$$

Multiply (17) through by $dl/\mathbf{B}$ where dl is path length along a line of force, and use (9) to obtain

$$\mathbf{dl}\cdot\mathbf{A}'=\mathbf{dl}\cdot\nabla\chi. \tag{18}$$

In another notation, this is

$$\mathbf{dl}\cdot\mathbf{A}'=d\chi. \tag{19}$$

Integrate this along a line of force from the reference surface to the point at which it is desired to compute $\mathbf{A}$. The result is

$$\chi=\chi_0+\int\mathbf{dl}\cdot\mathbf{A}', \tag{20}$$

Where χ_0 is obviously a function only of position on the reference surface, that is, if one uses α, β, V as the coordinates, χ_0 is a function only of α, β. But then

$$\mathbf{B}\cdot\nabla\chi_0=(\partial\chi_0/\partial\alpha)\mathbf{B}\cdot\nabla\alpha+(\partial\chi_0/\partial\beta)\mathbf{B}\cdot\nabla\beta=0$$

so that χ_0 satisfies (17) by itself, and we can set it equal to zero if we wish. Then

$$\chi=\int\mathbf{dl}\cdot\mathbf{A}'. \tag{21}$$

For computing $\mathbf{A}$, we need the gradient of this function, which we now obtain. Construct the following closed curve. Take that segment of a line of force which goes from the reference surface to the point at which the vector potential is to be computed. From this point proceed an infinitesimal distance in some direction. Denote this displacement by $\boldsymbol{\delta}_1\mathbf{s}$. Take that segment of a line of force which begins at this new point and ends on the reference surface. Close the curve with the segment lying in the reference surface and composed of points lying on lines of force which pass through $\boldsymbol{\delta}_1\mathbf{s}$. Call this new segment $\boldsymbol{\delta}_0\mathbf{s}$. This closed curve encloses no lines of force, so that if s denotes path length along it,

$$\oint\mathbf{A}'\cdot\mathbf{ds}=0. \tag{22}$$

Let d/ds denote the directional derivative in the direction $\boldsymbol{\delta}_1\mathbf{s}$. Then from (22) and the definition of a derivative it follows that

$$\frac{d}{ds}\int\mathbf{A}'\cdot\mathbf{dl}=\mathbf{A}'\cdot\boldsymbol{\delta}_1\mathbf{s}(\delta_1 s)^{-1}-\mathbf{A}_0'\cdot(d\mathbf{s}_0/ds)$$

where we define

$$d\mathbf{s}_0/ds=\lim_{\delta_1 s\to 0}(\boldsymbol{\delta}_0\mathbf{s}/\delta_1 s). \tag{23}$$

If we successively put $\boldsymbol{\delta}_1\mathbf{s}$ in the r, ϑ, and φ directions, we obtain expressions for the spherical components of $\nabla\chi$ which, when substituted in (16) yield

$$\begin{aligned}A_r&=\mathbf{A}_0'\cdot d\mathbf{s}_0/dr\\ A^\alpha&=r^{-1}\mathbf{A}_0'\cdot d\mathbf{s}_0/d\vartheta\\ A_\varphi&=(r\sin\vartheta)^{-1}\mathbf{A}_0'\cdot d\mathbf{s}_0/d\varphi.\end{aligned} \tag{24}$$

We have thus shown by construction that it is possible to adjust the gauge of $\mathbf{A}$ so that (3) is satisfied.

In order to use this result in computations it would be necessary to construct lines of force numerically.

CALCULATION OF α AND β

We first show that any function which is constant along every line of force may be chosen as α and that β is then determined.

It is obvious that there are exactly two independent functions both constant along every line of force. Clearly α and β are such a pair. Any other function which is constant along every line of force can be expressed as a function of α and β only. Suppose we have some arbitrarily given function $g(\alpha,\beta)$ and another function $h(\alpha,\beta)$ with properties to be determined. We now show that if h is defined in a certain way, g may be used in place of α provided h is used in place of β.

Any choice of α and β is appropriate provided (7) is satisfied. Then calculate

$$\nabla g\times\nabla h=(g_1\nabla\alpha+g_2\nabla\beta)\times(h_1\nabla\alpha+h_2\nabla\beta) \tag{25}$$

where the subscript denotes the obvious partial differentiation. We can rewrite (25) as

$$\nabla g\times\nabla h=(g_1h_2-g_2h_1)\nabla\alpha\times\nabla\beta$$

or

$$\nabla g\times\nabla h=(g_1h_2-g_2h_1)\mathbf{B} \tag{26}$$

by (7). Now, g and h can be used in place of α and β provided

$$g_1h_2-g_2h_1=1. \tag{27}$$

Since g is a known function, this is a partial differential equation for h which evidently possesses a solution. This therefore shows that α can be chosen arbitrarily and that β can then be computed.

In order to construct α and β, we consider the following way of specifying functions constant along lines of force. Suppose that B, considered as a function of position along a line of force, has a single minimum on each line of force. Call the surface composed of all the points at which such minima occur *the minimum surface* and denote quantities evaluated on it with a subscript m. Let there be chosen some arbitrary function $g(\mathbf{X}_m)$ of position on the surface. If the number $g(\mathbf{X}_m)$ be assigned to every point lying on the line of force passing through $\mathbf{X}_m$, the result is a function defined at every point in space which is constant along every line of force. We will denote such functions simply by $g(\mathbf{X}_m)$.

We now suppose α to be defined, using this method, and show how to compute β. First, define a second function γ which is constant along every line of force and independent of α. Put

$$\mathbf{B}=\nabla\alpha\times\nabla\beta(\gamma)=\beta'(\gamma)\nabla\alpha\times\nabla\gamma.$$

But then

$$\beta'(\gamma)=B/|\nabla\alpha\times\nabla\gamma|$$

so that

$$\beta=\int\frac{B\,d\gamma}{|\nabla\alpha\times\nabla\gamma|}. \tag{28}$$

The practical evaluation of this expression is of course quite tedious, requiring stepwise integrations along a large number of lines of force in order to evaluate the integrand at the same number of points along the path of integration. The integration path is, of course, the one for which α and V are both constant.

THE EQUATIONS OF MOTION

We now obtain the equations of motion for a charged particle moving in a magnetic field, using the functions α, β, and V as coordinates. The Lagrangian for this problem is well known to be

$$\mathfrak{L}=\tfrac{1}{2}mv^2+(e/c)\mathbf{v}\cdot\mathbf{A} \tag{29}$$

where $\mathbf{v}$ is the particle velocity, e its charge, m its mass, c the speed of light, and $\mathbf{A}$ the vector potential to any desired gauge. It is also well known that since v is a constant of the motion in the correct relativistic case, (29) is the correct relativistic Lagrangian provided m is taken to be a constant numerically equal to $m_0(1-v^2/c^2)^{-1/2}$ with m_0 the rest mass of the particle.

Let us first calculate $\mathbf{v}\cdot\mathbf{A}$ in our new coordinate system. We can evidently put

$$\mathbf{v}\cdot\mathbf{A}=\mathbf{A}\cdot d\mathbf{s}/dt \tag{30}$$

where $d\mathbf{s}$ is the vector displacement along a trajectory. Throughout what follows, take $\mathbf{A}$ to the gauge given by (3). Then

$$\mathbf{A}\cdot d\mathbf{s}=\alpha d\mathbf{s}\cdot\nabla\beta.$$

Assume that $\partial\beta/\partial t=0$, then from this last,

$$\mathbf{A}\cdot d\mathbf{s}=\alpha\,d\beta$$

so that (30) becomes

$$\mathbf{v}\cdot\mathbf{A}=\alpha\dot{\beta}. \tag{31}$$

Now we must transform $\frac{1}{2}mv^2$. In spherical coordinates,

$$v^2=r^2+r^2\dot{\vartheta}^2+r^2\sin^2\vartheta\dot{\varphi}^2. \tag{32}$$

Regarding each of r, ϑ, and φ as a function of α, β and V, carrying out the differentiations, rearranging the terms, and using (14) to drop those proportional to $\dot{\alpha}\dot{V}$ and $\dot{\beta}\dot{V}$, (32) becomes

$$v^2=\left[\left(\frac{\partial r}{\partial\alpha}\right)^2+r^2\left(\frac{\partial\vartheta}{\partial\alpha}\right)^2+r^2\sin^2\vartheta\left(\frac{\partial\varphi}{\partial\alpha}\right)^2\right]\dot{\alpha}^2+\left[\left(\frac{\partial r}{\partial\beta}\right)^2+r^2\left(\frac{\partial\vartheta}{\partial\beta}\right)^2+r^2\sin^2\vartheta\left(\frac{\partial\varphi}{\partial\beta}\right)^2\right]\dot{\beta}^2$$

$$+\left[\left(\frac{\partial r}{\partial V}\right)^2+r^2\left(\frac{\partial\vartheta}{\partial V}\right)^2+r^2\sin^2\vartheta\left(\frac{\partial\varphi}{\partial V}\right)^2\right]\dot{V}^2+2\left[\frac{\partial r}{\partial\alpha}\frac{\partial r}{\partial\beta}+r^2\frac{\partial\vartheta}{\partial\alpha}\frac{\partial\vartheta}{\partial\beta}+r^2\sin^2\vartheta\frac{\partial\varphi}{\partial\alpha}\frac{\partial\varphi}{\partial\beta}\right]\dot{\alpha}\dot{\beta}. \tag{33}$$

Upon computing the derivatives in the straightforward way, using the elementary theory of implicit functions, and substituting the results in (33), we obtain

$$v^2=\frac{A^2}{\alpha^2B^2}\dot{\alpha}^2+\frac{|\nabla\alpha|^2}{B^2}\dot{\beta}^2+\frac{1}{B^2}\dot{V}^2-2\frac{\mathbf{A}\cdot\nabla\alpha}{\alpha B^2}\dot{\alpha}\dot{\beta}. \tag{34}$$

The Langrangian is then

$$\mathfrak{L}=\tfrac{1}{2}m\left[\frac{A^2}{\alpha^2B^2}\dot{\alpha}^2+\frac{|\nabla\alpha|^2}{B^2}\dot{\beta}^2+\frac{1}{B^2}\dot{V}^2-2\frac{\mathbf{A}\cdot\nabla\alpha}{\alpha B^2}\dot{\alpha}\dot{\beta}\right]+(e/c)\alpha\dot{\beta}. \tag{35}$$

The equations of motion are then simply Lagrange's equations.

THE INTEGRAL

In the new coordinates, the part of the Lagrangian which depends on the magnetic field is independent of both β and V. It is obvious that the remaining part is strongly V dependent. For the existence of an integral, it is necessary that there be a group of transformations which leaves the Lagrangian invariant. The simplest case is when a particular coordinate is absent from the Lagrangian.

Let us then try the assumption that the Lagrangian is approximately independent of β. We have one of Lagrange's equations

$$(d/dt)(\partial\mathfrak{L}/\partial\dot{\beta})-(\partial\mathfrak{L}/\partial\beta)=0$$

and since $\partial\mathfrak{L}/\partial\beta=0$, we have the integral

$$\partial\mathfrak{L}/\partial\dot{\beta}=\text{constant}.$$

From (35) this becomes

$$m[(|\nabla\alpha|^2/B^2)\dot{\beta}-(\mathbf{A}\cdot\nabla\alpha/\alpha B^2)\dot{\alpha}]+(e/c)\alpha=\text{constant}. \tag{36}$$

Define the following angles.

$$\mathbf{v}\cdot\nabla\alpha=v\,|\,\nabla\alpha\,|\cos\eta$$
$$\mathbf{v}\cdot\nabla\beta=v\,|\,\nabla\beta\,|\cos\omega$$
$$\nabla\alpha\cdot\nabla\beta=|\,\nabla\alpha\,||\,\nabla\beta\,|\cos\zeta.$$

Denote by Φ the dihedral angle between the planes $\nabla\alpha$, $\nabla\beta$ and $\nabla\alpha$, $\mathbf{v}$. Set

$$\cos\psi=\sin\eta\cos\Phi.$$

Then

$$\dot{\beta}=\mathbf{v}\cdot\nabla\beta=v\,|\,\nabla\beta\,|\cos\omega$$
$$\dot{\alpha}=\mathbf{v}\cdot\nabla\alpha=v\,|\,\nabla\alpha\,|\cos\eta$$
$$\mathbf{A}\cdot\nabla\alpha=A\,|\,\nabla\alpha\,|\cos\zeta,$$

and (36) becomes

$$mv(A\,|\,\nabla\alpha\,|^2/\alpha B^2)(\cos\omega-\cos\eta\cos\zeta)+(e/c)\alpha=\text{constant}. \tag{37}$$

From the definitions and the spherical law of cosines it follows that

$$\cos\omega-\cos\eta\cos\zeta=\sin\eta\sin\zeta\cos\Phi. \tag{38}$$

Also,

$$B=|\,\nabla\alpha\times\nabla\beta\,|=|\,\nabla\alpha\,||\,\nabla\beta\,|\sin\zeta$$

so that

$$A\,|\,\nabla\alpha\,|\sin\zeta=\alpha B. \tag{39}$$

Using (38) and (39) together with the definition of ψ in (37), we obtain

$$mv(|\,\nabla\alpha\,|/B)\cos\psi+(e/c)\alpha=\text{constant}. \tag{40}$$

Assume that β is dimensionless. Then put (40) into Størmer units, that is, let $(M/R)^{-1/2}$ be the unit of length, where R is the particle rigidity and M is some constant appropriate to the magnetic field in question which has the units of dipole moment. We get

$$(|\,\nabla\alpha\,|/B)\cos\psi+\alpha=2\gamma$$

or

$$\cos\psi=(2\gamma-\alpha)B/|\,\nabla\alpha\,| \tag{41}$$

where our integration constant is now called 2γ. This is the result that we have been seeking.

The requirement that $\mathfrak{L}$ be approximately independent of β will not be investigated in detail here. If, in pursuing such an investigation one wished to compute $\partial\mathfrak{L}/\partial\beta$, it might be useful to note that, by the same procedures used in obtaining (34) one can show that

$$\partial f/\partial\beta=B^{-2}(\mathbf{B}\times\nabla\alpha)\cdot\nabla f. \tag{42}$$

It is obvious from (35) that requiring $\mathfrak{L}$ to be independent of β is equivalent to requiring that each of A^2, B^2, $|\nabla\alpha|^2$, and $\mathbf{A}\cdot\nabla\alpha$ be so.

AXIAL SYMMETRY

We now obtain (41) more explicitly for the case where the magnetic field has axial symmetry. This will show simply that this limiting case gives back the usual theory.

We have, obviously,

$$A_r = A_\vartheta = 0$$

$$\alpha = r\sin\vartheta A_\varphi$$

$$B = |\nabla\alpha||\nabla\beta|.$$

As a result,

$$B/|\nabla\alpha| = |\nabla\beta| = A_\varphi/\alpha.$$

Using the value of α, this yields

$$B/|\nabla\alpha| = (r\sin\vartheta)^{-1}. \tag{43}$$

Then (41) becomes

$$\cos\psi = [2\gamma/(r\sin\vartheta)] - A_\varphi. \tag{44}$$

Finally, from the definition of ψ and (38) together with the fact that now $\zeta = \pi/2$, we get

$$\cos\psi = \cos\omega,$$

so (44) becomes

$$\cos\omega = [2\gamma/(r\sin\vartheta)] - A_\varphi. \tag{45}$$

It is clear from its definition that this ω is the usual one, and (45) is the usual result when the magnetic field has axial symmetry.

TRAPPED RADIATION

Evidently, no trajectory can pass through any point where

$$|(2\gamma-\alpha)(B/|\nabla\alpha|)| > 1,$$

since it is then impossible to satisfy (41). Any such point will be called *forbidden*. Every other point will be said to be *allowed*.

It is evident that the line of force $\alpha = 2\gamma$ is contained in an allowed region, and that for reasonable magnetic fields there is some value of γ large enough that for all greater values of γ it is impossible to go from this allowed region to any point with $r = \infty$ along a path consisting entirely of allowed points. From this case we then study trapped radiation.

Let us put (41) into a different form. First a simple rearrangement gives

$$2\gamma = \alpha + (|\nabla\alpha|/B)\cos\psi. \tag{46}$$

We now do a succession of rearrangements of the right hand term. First, multiply its numerator and denominator each by vB and use the fact that B and $\nabla\alpha$ are perpendicular to each other to obtain

$$(|\nabla\alpha|/B)\cos\psi = (v|\mathbf{B}\times\nabla\alpha|\cos\psi)/vB^2.$$

From the definition of ψ it then follows from this that

$$\frac{|\nabla\alpha|}{B}\cos\psi = \frac{\mathbf{v}\cdot(\mathbf{B}\times\nabla\alpha)}{vB^2} = \frac{\nabla\alpha\cdot(\mathbf{v}\times\mathbf{B})}{vB^2}$$

where in the last step we simply take a cyclic permutation of the terms in the triple scalar product. With lengths in Størmer units, the larmor radius vector is

$$\mathbf{a} = (bB^2)^{-1}\mathbf{v}\times\mathbf{B}$$

where we take the vector to be directed from the particle toward the guiding center. Thus we obtain

$$(|\nabla\alpha|/B)\cos\psi = \mathbf{a}\cdot\nabla\alpha$$

so that (46) becomes

$$2\gamma = \alpha + \mathbf{a}\cdot\nabla\alpha. \tag{47}$$

Up to this point the result is exactly equivalent to (41). The right member of (47) is the sum of the first two terms of a Taylor expansion of the value of α at the guiding center about the position of the particle. If we denote the value of α at the guiding center by α_c and neglect terms of second order and above in the Taylor expansion, (47) becomes

$$2\gamma \cong \alpha_c. \tag{48}$$

This result clearly is independent of units.

The implication of (48) is that whenever (41) holds with good accuracy, there exists with the same accuracy a function of position which is constant along a line of force and is also a constant of the motion for trapped particles. These are the two properties that McIlwain's parameter L was constructed to possess approximately.[2] In fact, L was constructed to be an adiabatic invariant and shown numerically to be constant along lines of force to a percent or so. In our case, α_c is taken to be exactly constant along lines of force and shown under appropriate assumptions to be an approximate constant of the motion. We shall not worry about this distinction. These two functions have yet another important connection. For magnetic fields that become dipole fields at large distances, we have

$$\alpha = \sin^2\vartheta/r$$

$$L^{-1} = \sin^2\vartheta/r$$

at large distances. On the basis of these shared properties, we can put

$$\alpha_c = L^{-1} \tag{49}$$

throughout the remainder of this paper.

It is instructive to attempt to obtain (48) in another way, and to see why this attractive method does not work.

We return to Lagrange's equation which, after integrating with respect to time along a trajectory becomes

$$\frac{\partial \mathfrak{L}}{\partial \dot{\beta}} = \int_0^t \frac{\partial \mathfrak{L}}{\partial \beta}\, dt + \text{constant}.$$

Carrying out the differentiation in the left member, this evidently becomes

$$mv\left[\frac{|\nabla\alpha|^2}{B^2}\beta' - \frac{\mathbf{A}\cdot\nabla\alpha}{\alpha B^2}\alpha'\right] + (e/c)\alpha = \int_0^t \frac{\partial \mathfrak{L}}{\partial \beta}\, dt + \text{constant},$$

where primes denote differentiation with respect to path length along the trajectory. Using the same transformations that led to (41) except the conversion to Størmer units, this becomes

$$mv\frac{|\nabla\alpha|}{B}\cos\psi + (e/c)\alpha = \int_0^t \frac{\partial \mathfrak{L}}{\partial \beta}\, dt + \text{constant}. \tag{50}$$

Consider now a sequence of trapped particles, all having their mirror points at the same value of B and all having the same value of I where

$$I = v^{-1}\oint v_l\, dl,$$

l being path length along the line of force and v_l being the component of $\mathbf{v}$ along $\mathbf{B}$. As one passes along the sequence, let the speed monotonically decrease from some initial speed of interest down to zero. Evidently all of the particles travel along the same magnetic shell. One can attempt to compute a label for that shell by taking the limit of (50) as v goes to zero. Obviously, the left member goes, in that limit, to $(e/c)\alpha_c$ so that

$$(e/c)\alpha_c = \lim_{v\to 0}\int_0^t \frac{\partial \mathfrak{L}}{\partial \beta}\, dt + \text{constant}. \tag{51}$$

Now, only the kinetic energy term of $\mathfrak{L}$ contributes to the integrand. We can then put

---

$$\frac{\partial \mathfrak{L}}{\partial \beta} = \tfrac{1}{2}mv^2\frac{\partial}{\partial\beta}\left[\frac{A^2}{\alpha^2B^2}\alpha'^2 + \frac{|\nabla\alpha|^2}{B^2}\beta'^2 + \frac{1}{B^2}V'^2 - 2\frac{\mathbf{A}\cdot\nabla\alpha}{\alpha B^2}\alpha'\beta'\right] \equiv \tfrac{1}{2}mv^2(\partial F/\partial\beta)$$

so that (51) becomes

$$(e/c)\alpha_c = \lim_{v\to 0}\tfrac{1}{2}mv^2\int_0^t \frac{\partial F}{\partial \beta}\, dt + \text{constant}. \tag{52}$$

Now, if every trajectory in the sequence were to be run for the same length of time T, this would lead to the conclusion that α_c was constant and no assumption about the smallness of $\partial F/\partial\beta$ would be necessary. However, what is really wanted is to use this result for a particle of vanishing v to compute a magnetic shell, and then use the fact that all of the particles in the sequence travel on the same shell to describe the motion of particles with nonvanishing v. Suppose there is a particle with speed v which one wishes to follow for a time T. Then it is necessary that the particle with vanishing v be followed long enough that its drift velocity carry it as far as the faster particle goes in time T. It is evident that the length of time for which a particle is followed must then be inversely proportional to its drift velocity. Then (52) becomes

$$(e/c)\alpha_c = \lim_{v\to 0} \tfrac{1}{2}mv^2 \int_0^{\tau/(1/2)mv^2} \frac{\partial F}{\partial \beta}\, dt + \text{constant}$$

or

$$(e/c)\alpha_c = \lim_{v\to 0} \int_0^{\tau} \frac{\partial F}{\partial \beta}\, dx + \text{constant} \qquad (53)$$

since drift velocity is proportional to $\frac{1}{2}mv^2$ and where we have put $x = \frac{1}{2}mv^2 t$. Since the integral is now independent of v, we have

$$(e/c)\alpha_c = \int_0^{\tau} \frac{\partial F}{\partial \beta}\, dx + \text{constant}. \qquad (54)$$

It is therefore evident that a condition on the accuracy of (41) is that

$$\int_0^{\tau} \frac{\partial F}{\partial \beta}\, dx \ll (e/c)\alpha_c. \qquad (55)$$

This is independent of v, adding to our confidence that (41) is not only good for low rigidities.

COSMIC RAY CUTOFFS

We now obtain some results concerning Størmer cutoffs in our magnetic field. Assume that we have the usual qualitative features present in the Størmer plots, that is, inner and outer allowed regions with a pass open between them provided γ is less than a certain critical value. Then, from (41), the curve $\alpha = 2\gamma$ lies in the inner allowed region. Let us suppose that at sufficiently large distances the field becomes that of a dipole, and that the line of force in question begins at the earth at sufficiently high latitude that part of it lies in this dipole region. Then from (44) we obtain the usual critical value of γ above which particles cannot arrive at the earth as $\gamma_c = 1$. Assuming that the latitude at which the particle arrives at the earth is sufficiently high that the approximate form of (41) given by (48) is valid, we then have $\alpha_c = 2$. On converting this from Størmer units back to a set normally used in treating data, this becomes $R = 14.9\alpha_c^2$ Bv where α_c^{-1} is in earth radii and is computed at the guiding center. Using (49) this becomes

$$R = 14.9L^{-2}\ \text{Bv}. \qquad (56)$$

The value 14.9 is of course obtained using the earth's dipole moment, and L must be computed at the guiding center of the particle. This result is the same as that obtained previously[3] using, among other things, Alfvén's description of the trajectory when it is near the earth.

The result embodied in (56) is only good at latitudes sufficiently high that the line of force enters a dipole region of the magnetic field. We now obtain a result which is good everywhere. It is clear that in the general case the critical value of γ is a function only of rigidity. Call this critical value γ_c. Let us now suppose that a set of observers are all stationed along the same line of force and that each is looking in some direction such that $\cos\psi = 0$. For each of these observers, from (41), $\alpha = 2\gamma$. The quantities in this equation are, of course, in Størmer units. All of these observers have the same value of α, when α^{-1} is measured in earth radii. If they all observe at the same value of R, they will then have the same value of α in Størmer units, and since $\alpha = 2\gamma$, they will have the same value of γ. If this value of R is the cutoff rigidity for one of these observers, it will then be so for all of them since the critical value of γ is a function only of R. We thus have the result *the cutoff at any point on a particular line of force in any direction for which* $\cos\psi = 0$ *is the same as that at any other point on the same line in any direction satisfying the same condition.*

In the particular case of the earth's field, we can get, with reasonably good accuracy, an actual

value for the cutoff in this particular direction. Suppose that the cutoff rigidity is less than 14.9 Bv. Then if the field were that of a dipole, the pass point between the outer and inner allowed regions would be farther than two earth radii from the center of the earth. In fact, the Finch and Leaton field beyond two earth radii is not very different from that of a dipole. Consequently, for cutoffs less than 14.9 Bv, we can take the critical value of γ to be unity, making errors in the cutoff value of less than twenty or thirty percent. [See the discussion of errors in[3].] At low latitudes we of course cannot use (48), but must use (41) instead. For a general direction this can be worked out in a straightforward way, but is evidently tedious. In the direction $\cos\psi=0$, we have $\alpha=2\gamma$. Then we have the result, using (49), that *when* $\cos\psi=0$ *and the cutoff so computed is less than* 14.9 Bv, *the cutoff is given by* (56) *with* L *computed at the observation point.* This has previously been verified[3] numerically for a set of lines of force in the Finch and Leaton field. From its definition, ψ is the angle between $\mathbf{v}$ and $\mathbf{B}\times\nabla\alpha$. Consequently $\cos\psi=0$ specifies the plane of directions which contains both $\mathbf{B}$ and $\nabla\alpha$. In particular, the direction along the line of force satisfies the condition.

We can compute the cutoff in any other direction in the following way. By putting $\gamma=1$ into (41) and then restoring cgs units, we obtain

$$\cos\psi=[2-\alpha(M/R)^{1/2}][(B/M)/|\nabla\alpha|](M/R)^{1/2}.$$

On solving for R, this becomes

$$R=M\{\alpha^{-1}\pm[\alpha^{-2}-\alpha^{-1}(B/M)^{-1}\mid\nabla\alpha\mid\cos\psi]^{1/2}\}^{-2}.$$

Convert lengths to earth radii and rigidity to Bv. We finally have

$$R=59.6\{\alpha^{-1}+[\alpha^{-2}-(\alpha B/M)^{-1}\mid\nabla\alpha\mid\cos\psi]^{1/2}\}^{-2}, \tag{57}$$

where, as usual, the positive sign is chosen for the square root so that the observation point will lie in the inner allowed region. As already noted, the errors in this expression will become greater than some twenty or thirty per cent for rigidities greater than 15 Bv.

IMPACT ZONES

A complete theory of impact zones requires the use of a computer already in the case of a dipole field. Webber[4] has found a way to make the impact zone calculations, for latitudes greater than some reasonable limit in a field which at sufficiently great distances is that of a dipole, depend on already known dipole results. He calls attention to the trajectories which, in a dipole field, pass through the dipole. There is one of these for each value of γ, and Størmer has calculated an extensive set of them.[5] Webber notices the following fact. Consider a particle arriving at the earth at reasonably high latitude. Then the last part of its path will be accurately represented by Alfvén's description, and the particle will pass through a large range of values of B while this is so. Because of Alfvén's relation

$$\sin^2\sigma/B=\text{constant}$$

with σ the angle between $\mathbf{v}$ and $\mathbf{B}$, this particle will be travelling nearly along a line of force far out from the earth where the field is that of a dipole. But then it will there be nearly indistinguishable from that trajectory through the dipole which belongs to the value of γ in question. Since these are already calculated for a dipole field, one can then simply calculate the value of γ for the particle in question and look up from Størmer's work the direction from which the particle is coming when it is at an infinite distance from the earth. Webber uses a heuristic argument based on the cutoff theory that he, Quenby, and Wenk developed in order to calculate γ for the observation point and direction at the earth's surface.

In the context of the present paper, working out impact zones from first principles requires the computation of trajectories, and (41) is of use only in reducing the order of the equations by 2. However, one can also adopt all of Webber's scheme for high latitudes and make the single change of computing γ using (41), or better, its approximate version (48). Then

$$\gamma=\alpha_c/2.$$

This is, from (49),

$$\gamma=(2L)^{-1} \tag{58}$$

with L computed at the guiding center and both α_c and L in Størmer units. Changing the units, (58) becomes

$$\gamma^2 = 14.9/(RL^2) \tag{59}$$

with R in Bv and L in earth radii, for the earth's dipole moment. As before, L is computed at the guiding center of the particle when it arrives at the earth, and the resulting value of γ, following Webber, is used to look up the appropriate trajectory through the dipole as computed by Størmer.

CAVITY FIELDS

We have obtained confidence in the accuracy of (41) by using it to obtain a variety of previously verified results. McIlwain's L parameter has played a central role in the argument, and an essential point has been that L is both an adiabatic invariant and nearly constant along field lines. Now, L is defined to be an adiabatic invariant, but it is known to be constant along field lines only through McIlwain's indefatigable use of an automatic computer and a particular magnetic field model. We have no assurance that L defined in the same way would still be constant along fields lines in such a greatly different field as, for example, that in the cavity formed by a solar wind blowing past the earth. An alternative definition has been suggested by Stone[6] which is on a different logical footing but in fields where such parameters exist has the same practical consequences. This proposed new definition has the added advantage that if it is verified not to have both properties to good accuracy, it is then certain that there is no function which does.

The definition consists in narrowing the definition of α in the paragraphs preceding Eq. (28) so that if there exists a choice of α which is an adiabatic invariant, the one selected will be so. Recalling the notation of those paragraphs, it is merely necessary to choose α to be a function of B_m only in order to insure this result, since the particle drifting with its mirror point magnetic field equal to B_m clearly drifts so that the magnetic field at its guiding center is constant. In any magnetic field where a function exists with both properties and where B_m as a function of position over its defining surface is that of a dipole field, by choosing $\alpha = (B_m/M)^{1/3}$ where M is the dipole moment, we then have $L = \alpha^{-1}$. If the accuracy of the invariance so produced be marginal, it can be somewhat improved by choosing the value of I for suitably chosen mirror points as the numerical value of α to assign to a given line of force.

All of the results previously obtained by various methods have been for fields with internal sources only, and for fields which have only somewhat small departures from axial symmetry except very near the earth. At the same time it would be very desirable to apply (41) to situations which are nowhere close to axial symmetry, such as the field inside a cavity produced by a solar wind of plasma blowing by the earth. We have no general discussion which demonstrates the validity of (41) under such circumstances, and for the time being must be content with some qualitative considerations that render reasonable a hope that this validity can be demonstrated.

In the first place, let us see what we can discover about the features of the earth's internal field which result in L being both conserved and nearly constant along a line of force.

From Alfvén's work[7] we know that in a dipole field the range of longitude through which a trapped particle drifts while it is spiraling from one mirror point to the next is nearly independent of the latitude of the mirror point for latitudes all the way from the equator to 60°. This clearly comes to be in the following way. The drift velocity, when $\nabla \times \mathbf{B} = 0$, is given by

$$\mathbf{v}_D = \kappa[(2/B) - B_R^{-1}](\mathbf{B}/B) \times (B^{-1}\nabla B),$$

where κ is some constant and B_R is the magnetic field at the mirror point. At the equator, $\mathbf{B}$ and ∇B are perpendicular to each other, but as one goes along the line of force away from the equator $\mathbf{B}$ begins to rotate fairly rapidly in the meridian plane while ∇B does not change its direction very much, and the angle between $\mathbf{B}$ and ∇B decreases, causing $|\,(\mathbf{B}/B) \times (rB^{-1}\nabla B)\,|$ to decrease. At the same time, B begins to increase, causing $(2 - B/B_R)$ to decrease. There remains a factor $(rB)^{-1}$ which decreases. As a result, as one goes away from the equator the drift velocity rapidly decreases. Since most of the drifting occurs while the particle is near the equator, the amount of

drift in longitude during one so called bounce period is nearly independent of the latitude range of the bounce. If one now alters the field by the addition of other terms which destroy the axial symmetry but don't alter the general structure of the field very much, so that each line of force still has a greatest distance from the earth and comes down to it on both ends, and so that, furthermore, no extremely large change is made in the rates at which $\mathbf{v}_D$ varies as one goes along a line of force, one can with reasonable confidence conclude that the rate of drift per bounce period should not depend importantly on the length of the bounce. More than this one can even expect that the average direction of drift may not change very much as the mirror point moves down the line of force. If it should turn out that this is so, then clearly all the particles initially spiraling about one line of force should travel nearly along the same magnetic shell. This makes it not unreasonable to hope that a function exists which is both constant along a line of force and approximately conserved during the motion of a trapped particle. This is a necessary, but of course not sufficient, condition for (41).

REFERENCES

1. GRAD, H., private communication to C. S. Gardner; *Phys. Rev.*, **115,** 791 (1959).
2. MCILWAIN, C. E., *J. Geophys. Research*, **66,** 3681 (1961).
3. SAUER, H., and RAY, E. C., *Ann. Phys.*, in press.
4. WEBBER, W. R., *J. Geophys. Research*, **8,** 3065 (1963).
5. STØRMER, C., "The Polar Aurora," Chap. VI. Oxford Univ. Press, London, 1955.
6. STONE, E. C., Univ. of Chicago Research Report EFIN, S–62–72 (1962), to be published.
7. ALFVÉN, H., "Cosmical Electrodynamics," p. 29, figure 2.4. Oxford Univ. Press, London, 1950.

THEOREMS CONCERNING THE ALLOWED CONE OF COSMIC RADIATION

ERNEST C. RAY*

Goddard Space Flight Center

A previous investigation of the cosmic ray shadow cone and low latitude main cone is extended to include the main cone at all higher latitudes. The results are embodied in two theorems which specify the nature of a boundary of a class of trajectories, none of which are forbidden. No numerical results are included.

INTRODUCTION

The present paper extends previous work[1] on the theorems underlying the work of Lemaitre and Vallarta[2] on cosmic ray cutoffs in a dipole magnetic field. In the previous paper, it was shown that for particles with rigidity greater than that of of Schremp's F_0 cutoff, but sufficiently far below that for which the earth's surface interects the outer periodic orbit, trajectories asymptotic to this periodic orbit form the boundary of a continuous set of trajectories, all of which leave the earth never to return. It was also shown that, for the same range of rigidities, all trajectories which leave the earth but are not in this allowed set subsequently intersect the earth. Finally, it was shown that Schremp's shadow trajectories bound a set trajectories, all of which leave the earth only to intersect it again.

The proofs of these theorems depend on the representation of the trajectories by curves in a certain three-dimensional phase space. There is a three-dimensional subspace of this space, through every interior point of which pass exactly two trajectories and through every boundary point of which passes exactly one trajectory. The complement of this subspace is devoid of real trajectories.

The process of formulating theorems in this representation is somewhat complicated by the passage of two trajectories through each point. In the present paper, we introduce a representation which aviods this inconvenience and which still can be readily visualized. This new representation is then exploited in the formulation and proof of certain theorems concerned with the main cone of cosmic radiation.

THE REPRESENTATION OF TRAJECTORIES

As discussed in ref. 1, we are concerned with properties of solutions of the system

$$dx/d\sigma = \pm f^{1/2}(x, \lambda, \lambda'; \gamma) \tag{1.a}$$

$$d\lambda'/d\sigma = g(x, \lambda) \tag{1.b}$$

$$d\lambda/d\sigma = \lambda' \tag{1.c}$$

where

$$f(x, \lambda, \lambda'; \gamma) = (2\gamma)^{-4}e^{2x} + 2e^{-x} - e^{-2x}\cos^2\lambda - \cos^{-2}\lambda - \lambda'^2$$

and

$$g(x, \lambda) = e^{-2x}\sin\lambda\cos\lambda - \sin\lambda\cos^{-3}\lambda.$$

All of the variables and constants are required to be real, and

$$-\pi/2 \leqq \lambda \leqq \pi/2, \qquad -\infty \leqq x \leqq \infty.$$

The details of the connection of these equations with the motion of an electrically-charged particle in a dipole magnetic field are given in reference 1 and reference 2.

*National Academy of Sciences fellow on leave from the Department of Physics and Astronomy, University of Iowa, Iowa City, Iowa.

It is evident from (1.a) and the reality of x and σ that trajectories occur only in that region of space for which

$$f(x, \lambda, \lambda'; \gamma) \geqq 0. \tag{2}$$

We will study sets of trajectories for all members of each of which γ has one fixed value. We will be interested only in cases where $\gamma^* \leqq \gamma \leqq 1$. The limiting γ^* is that value of γ below which the principle periodic orbits do not exist. Godart[3] has obtained $\gamma^* = 0.788541$. For any fixed value of γ, we will call the surface $f(x, \lambda, \lambda'; \gamma) = 0$ *the motion boundary*. As stated in (1), with $0 < \gamma \leqq 1$, the motion boundary has the shape of a "Y" constructed of hollow tubing. It is evident that its intersection with the plane $\lambda' = 0$ is the boundary of Størmer's forbidden region. When $0 < \gamma \leqq 1$, the Størmer boundaries have structure shown in Fig. 1. The shaded regions are forbidden. As shown in reference 1, exactly two trajectories pass through every allowed point not on the motion boundary, and exactly one passes through each point on that boundary. At every point not on the motion boundary, for one of the trajectories $dx/d\sigma > 0$, while for the other $dx/d\sigma < 0$.

We now introduce a representation of the trajectories which has the property that exactly one trajectory passes through every allowed point. Consider two separate three-dimensional phase spaces, each being in the variables (x, λ, λ'). Call one of these the *positive space*. Call the other the *negative* one. In each of these, construct the motion boundary $f(x, \lambda, \lambda'; \gamma) = 0$. Discard from each the entire forbidden region. Thus far, we have two copies of the same thing. We will say that two points *correspond* when one belongs to one of these spaces and the other to the other, provided that $x_1 = x_2$, $\lambda_1 = \lambda_2$, and $\lambda'_1 = \lambda'_2$ where $(x_1, \lambda_1, \lambda'_1)$ represents one of the points and $(x_2, \lambda_2, \lambda'_2)$ represents the other. Notice that we are not regarding as identical the members of every such pair of points. Now identify the two motion boundaries in the obvious way. That is, consider to be a single point any two corresponding points, each of which lies in the motion boundary of its own space. The result is a single new space. We will continue to speak of the positive (or negative) space, by which we will now mean that part of the new space which was originally a subspace of the positive (or negative) space.

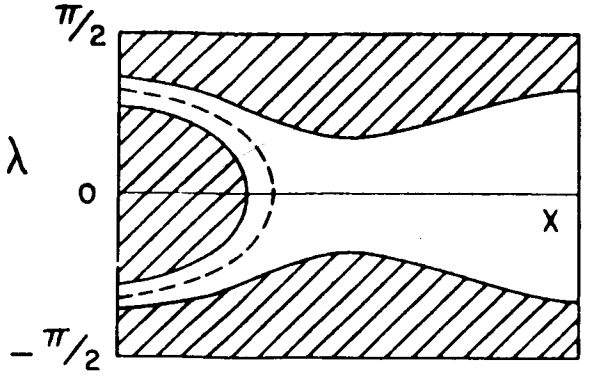


FIGURE 1.—A Størmer plot when $0 < \gamma < 1$. The shaded regions are forbidden.

The trajectories are now to be plotted in the following way. Each trajectory is composed of segments everywhere on each of which $dx/d\sigma$ has a constant sign. The segments are all bounded by points at which $dx/d\sigma = 0$. These bounding points all lie in the motion boundary and the motion boundary consists entirely of such points. Any segment for which $dx/d\sigma > 0$ is now plotted in the positive space, while any segment for which $dx/d\sigma < 0$ is plotted in the negative space. A trajectory is clearly continuous in this representation, and it moves back and forth from one space to the other by passing through the motion boundary. It is obvious that through every point in the space passes exactly one trajectory.

Any fixed plane $x = x_e$, where x_e is a constant, will be called the *earth's surface*. Any point for which $x < x_e$ will be said to lie *below* the earth's surface.

Any half trajectory which begins at a point in the earth's surface, has no points below the earth's surface, is not asymptotic to an orbit other than the outer periodic orbit, and is not periodic, will be said to be *allowed*. A trajectory which begins at the earth's surface in the positive space and subsequently enters the earth's surface in the negative space will be said to be *forbidden*. An allowed half trajectory may have any number of isolated points lying in the earth's surface.

THE ALLOWED CONE

We now formulate and prove the theorem which is fundamental to the theory of the allowed cone of cosmic radiation in a dipole field.

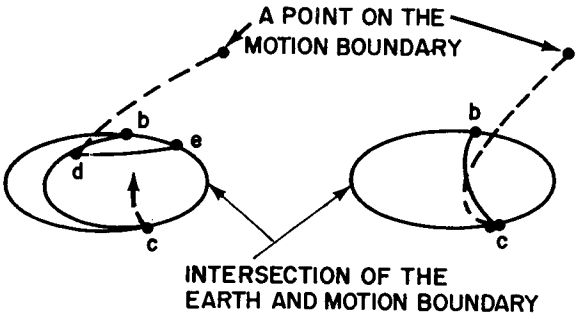


FIGURE 2.—The initial curve when none of it lies in the motion boundary but part of it lies in the negative space. The three dashed arcs terminated as arrows together constitute the trajectory which goes from d to a minimum at c. (See the text.)

Consider a simple closed curve which lies partly in that part of earth's surface which is in the positive part of the space. It may or may not have segments which lie in the motion boundary above the earth's surface or in that part of the earth's surface which is in the negative space. In the first of these optional cases, every point of it which lies in the motion boundary and every point of that part of the motion boundary bounded partly by it and partly by a curve in the earth's surface must be such that $\partial f/\partial x>0$. Here, f is the function defined just after Eq. (1). In second of the optional cases, we assume the following rather special structure (see Fig. 2). Assume there are exactly two intersections of the initial curve with the motion boundary. Call one b and the other c. Suppose there is a point d in the initial curve which lies in the earth's surface in the positive space and from which issues a trajectory which becomes tangent to the earth's surface at c, having there a minimum in x. Suppose that there is a segment in the earth's surface from d to a point e in the motion boundary where e lies between b and c in that region bounded by the closed curve we started with, and that this segment is such that every trajectory which issues from it has a tangency with the earth's surface at a point on the motion boundary between e and c, where, as the initial point moves from d to e, the tangency point moves from c to e. Now, delete from the simple closed curve mentioned at the beginning of this paragraph the arc dbc and add the arc dec. The result is again a simple closed curve, none of which lies in the negative space. Call the resulting curve the *initial curve*.

Constant the following surface, to be called the *trajectory surface*. Take every half trajectory which begins on any part of the initial curve other than an arc of the type of dec. Take that segment of each trajectory which begins on de and ends on ec. If every trajectory which does not end on ec is allowed, we will say that the trajectory surface is allowed.

Call that part of the earth's surface which is bounded (partly) by the initial curve (and partly by a part of the motion boundary where $\partial f/\partial x>0$) *the initial part* of the earth's surface.

When part or all of the trajectories in the trajectory surface area symptotic to the outer periodic orbit, we also include all or an obvious part of those trajectories which both have $dx/d\sigma>0$ almost everywhere and are asymptotic to the same trajectory in the infinite past.

THEOREM 1. *Assume there is an initial curve for which the associated trajectory surface is allowed. No half trajectory which begins in the initial part of the earth's surface is forbidden.*

As shown in Appendix I, there are no curves of intersection of the trajectory surface and the motion boundary which cannot be shrunk in the motion boundary to a point. Any such curve of intersection which can also be shrunk in the trajectory surface to a point obviously makes available to trajectories which start in the initial part of the earth's surface no access route to that part of the earth's surface which lies in the negative space. If there is a curve of intersection which cannot be shrunk in the trajectory surface to a point, it must, since the trajectories are allowed, be followed by another with the same property, and again the negative part of the earth's surface does not become accessable.

It is so that

$$d^2x/d\sigma^2=\tfrac{1}{2}\partial f/\partial x. \tag{3}$$

Consequently, every trajectory which passes through the motion boundary at a point where $\partial f/\partial x>0$ has there a minimum in x, and is thus going from the negative to the positive space. Consequently, no trajectory which starts in the initial part of the earth's surface can penetrate

the motion boundary in those regions not shielded by trajectories.

Finally, it is clear that the special structure assumed in connection with the second optional property of the initial curve supplies an impenetrable cap which shields the initial part of the earth's surface from that part of the earth's surface which lies in the negative space.

The theorem is established.

We now consider the set of trajectories which begin in initial part of the earth's surface and are not allowed. By Theorem I and the definitions of allowed and forbidden trajectories, the only trajectories in this set are either both periodic and tangent to the earth's surface, or asymptotic to a trajectory other than the outer periodic orbit. The first of these does not exist. A periodic orbit which was tangent to the earth's surface would have to pass through that part of the motion boundary which is partly bounded by the initial curve. At some preceding point, it would have passed through a part of the motion boundary where $\partial f/\partial x<0$. It is clear from the proof of Theorem 1 that there is no path which lies in the negative space and joins two such regions of the motion boundary, where the region with $\partial f/\partial x<0$ is accessable to trajectories starting in the earth's surface.

This leaves us only the trajectories asymptotic to trajectories other than the outer periodic orbit. It is an obvious corollary of a theorem of Hopf that these trajectories are a set of zero measure on the set of all trajectories which begin in the initial part of the earth's surface. We quote Hopf's Theorem I as given by Nemytskii and Stepanov[4]:

Let there be given a locally compact metric space R of motions with a countable base. Let there be defined in it an invariant measure μ having the following properties: $\mu R=+\infty$, but for any compact set $F\subset R$ the measure μF is finite. Then almost all points $p\in R$ as $t\to+\infty$ are either stable according to Poisson or are departing.

We refer to reference 4 for a detailed discussion of this theorem. It is obvious that all the hypotheses of the theorem are satisfied. (See Appendix II. The characteristic property of a trajectory which is stable according to Poisson is that infinitely often in the future it passes arbitrarily near any initial point of itself. In our case, such a trajectory is obviously forbidden, and by our Theorem 1, above, there are no trajectories of this sort which begin in the initial part of the earth's surface. A departing point is one such that the trajectory through it has no ω limit points. The trajectory to which an asymptotic trajectory is asymptotic consists, however, of ω limit points of the asymptotic trajectory.[5] We then infer from our own definitions, our Theorem 1, and Hopf's Theorem, the following:

THEOREM 2. *Zero is the measure of the set of those trajectories which begin in the initial part of the earth's surface and are not allowed.*

Since, as is obvious every trajectory which is allowed is either unbounded or asymptotic to the outer periodic orbit, this pair of theorems establishes all the properties of trajectories desired for the theory of the main cone cosmic radiation. This is the case when allowed trajectories in the trajectory surface are all asymptotic to the outer periodic orbit.

Appendix I

Here we demonstrate that there are no curves of intersection of the trajectory surface with the motion boundary which cannot be shrunk, in the motion boundary, to a point.

Suppose there were such a curve. Then it cannot consist entirely of trajectory minima and inflection points. We show this by considering the regions on the motion boundary where $dx^2/d\sigma^2$ has a particular sign. We have

$$d^2x/d\sigma^2 \equiv x'' = ae^{2x} - e^{-x} + e^{-2x}\cos^2\lambda \qquad (4)$$

where $a \equiv (2\gamma)^{-4}$. The intersection of the motion boundary with the plane $\lambda' = 0$ is given by

$$ae^{2x} + 2e^{-x} - e^{-2x}\cos^2\lambda - \cos^{-2}\lambda = 0. \qquad (5)$$

On solving (5) for $\cos^2\lambda$ and subsituting in (4), we obtain

$$x''_{\pm} = (3a/2)e^{2x} \pm [(a^2/4)e^{4x} + ae^{x}]^{1/2} \qquad (6)$$

where, for the plus sign $\cos^2\lambda > e^x$ and for the minus sign $\cos^2\lambda < e^x$. The curve $\cos^2\lambda = e^x$ is shown in Fig. 1 as a dashed curve. Consequently, on the Størmer boundary, when $\cos^2\lambda > e^x$, $x'' > 0$, and trajectories have only minima. To investigate the sign of x''_-, we note that it evidently has the sign as Q where

$$Q \equiv a^{-1}x''_+ x''_- e^{-x}\gamma.$$

Then

$$Q = r^3 - \gamma$$

where $x \equiv \ln 2\gamma r$ and physically, r is the distance from the dipole to the point in question. Consequently, for that part of the Størmer boundary where $\cos^2 \gamma < e^x$, trajectories have minima when $r > \gamma^{1/3}$ and maxima when $r < \gamma^{1/3}$. This critical value of r is large enough that the restricted main cone theorem[1] rules out a curve of intersection which cannot, in the motion boundary, be shrunk to zero. From these considerations, it is clear that no such curve can consist only of minima.

Next, we show that no such curve can consist partly of minima and partly of maxima. Suppose it consisted of one open segment of minima, one open segment of maxima, and two inflection points. Then just preceding each minimum, the same trajectory must have a maximum. (We obviously exclude from consideration the case where part of the curve consists of part of the motion boundary.) It is obvious that these preceding maxima, the two inflection points, and the locus of minima make up a closed curve which can be shrunk, both in the motion boundary and the trajectory surface. This case is then covered in the main body of the proof.

Suppose, finally, there were such a curve consisting entirely of maxima (and, perhaps, inflection points). Since all of the trajectories are allowed, there must be a later curve of intersection which consists entirely of minima, and which consequently can be shrunk in the motion boundary to a point. The segment of the trajectory surface bounded by these two curves of intersection would, by the uniqueness theorem, prevent the curve $\lambda = 0$, $\lambda' = 0$ from being a trajectory. It is clear from the equations of motion that this curve is a trajectory. It is now clear that the trajectory surface and motion boundary can have no curve of intersection which cannot be shrunk, in the motion boundary, to a point.

Appendix II

We exhibit here the invariant measure required by Hopf's theorem. It is the integral invariant that is associated with Jacobi's multiplier. If and only if

$$\int M(x, \lambda, \lambda')\, dx\, d\lambda\, d\lambda'$$

is an integral invariant,

$$\pm(\partial/\partial x)(Mf^{1/2})+(\partial/\partial\lambda)(M\lambda')+(\partial/\partial\lambda')(Mg)=0$$

Choose $M=f^{-1/2}$ and notice that $g=\frac{1}{2}(\partial f/\partial\lambda)$. Then this last condition becomes

$$\lambda'(\partial/\partial\lambda)(f^{-1/2})+\tfrac{1}{2}(\partial/\partial\lambda)\cdot(\partial/\partial\lambda')(f^{-1/2})=0$$

Using the definition of f, the left member vanishes so that this equation in an identity, as required.

REFERENCES

1. Ray, E. C., and Kasper, J. E., *Ann. Phys. (N.Y.)*, **20,** 119 (1962).
2. Vallarta, M. S., "An Outline of the Allowed Cone of Cosmic Radiation," Univ. of Toronto Press, Toronto, 1938.
3. Godart, O., *Ann. Soc. Sci. Bruxelles*, **A58,** 27 (1938).
4. Nemytskii, V. V., and Stepanov, V. V., "Qualitative Theory of Differential Equations," p. 454, Princeton Univ. Press, Princeton, New Jersey, 1960.
5. ———, and ———, *op. cit.*, pp. 401, 338.

ON COSMIC RAY CUTOFFS

HERBERT H. SAUER
State University of Iowa
AND
ERNEST C. RAY*
Goddard Space Flight Center

A cosmic ray cutoff in the Størmer sense is defined for sufficiently high latitudes when the magnetic field, while not axially symmetric, becomes approximately so at sufficiently great distances from the earth. It is shown that the cutoff, when so defined, is given by a simple function of McIlwain's L parameter. Numerical comparisons are made with a previous computer study of cutoffs defined in the same way. The expression for cutoff in terms of L suggests that the cutoff in the direction of the magnetic field should be approximately constant along a line of force. This is shown numerically to be so at latitudes too low for the derivation of the L expression to hold, for a particular field model.

INTRODUCTION

One of us[1] has conducted a numerical study of cosmic ray cutoffs at latitudes above 45° geomagnetic field which beyond a certain distance is approximately axially symmetric. The computations were done by assuming that Alfvén's[2] description of the motion of a charged particle in a magnetic field is valid when the particle is near the earth, and that it remains so until the particle is so far from the earth that the field is approximately axially symmetric and a Størmer integral exists. The Alfvén method describes the motion as a spiral about a line of force with a pitch angle which varies in a given way and with an instantaneous center which slowly drifts from one line to force to another as the motion proceeds. This drift was neglected. The cutoff computation depended on a numerical quadrature to obtain the phase in its spiraling of the particle at the point where it was finally assumed to be traveling in a field with axial symmetry. This phase made possible the calculation of its direction of motion so that the Størmer integral could be used to see whether or not it was forbidden by the Størmer criterion. It was also shown that the intuitively appealing expression

$$R = 14.9L^{-2}\mathrm{Bv} \qquad (1)$$

agrees with the results obtained to a difference of some 30%.

In the present paper, we show that (1) holds with very small errors (a percent or so) provided that L is calculated at the guiding center rather than the observation point, and provided the trajectory as described by the Alfvén method goes far enough from the earth to enter a region beyond which the field is approximately axially symmetric.

However, only part of the previous difference between (1) and the numerically computed values arises from computing L in reference 1 at the observation point. The rest is an error in the numerically obtained result which arose from applying the Størmer integral at a radius so small that some of the nondipole terms, while contributing less than one per cent to the magnetic field, significantly affected the motion.

It is one obvious consequence of (1) that the

*National Academy of Sciences fellow on leave from the Department of Physics and Astronomy, State University of Iowa, Iowa City, Iowa. Present address: Center for Radiophysics and Space Research, Cornell University, Ithaca, New York.

cutoff in the direction of the magnetic field is approximately constant alone a line of force. We numerically verify that this remains so at latitudes lower than those for which (1) holds, by integrating numerically the exact equations of motion for the Finch and Leaton field.[3]

It is, of course, natural to conjecture that (1) holds at all latitudes with good accuracy. We have, as yet, no way of showing that, other than by comparing it with numerical integrations, and we have not done this.

THE STØRMER CONE

We will continue to speak of the Størmer cone when referring to (1).

We can lay a useful intuitive background in the following way. Figure 1 shows a well known Størmer plot[4] for a dipole field when γ (Størmer's γ_1) is below, but near, unity. In the shaded region, motion is impossible because of the Størmer integral. The dashed curve is the only line of force which is entirely contained in an allowed region. Its equation is

$$r^{-1}\sin^2\vartheta = 2\gamma \qquad (2)$$

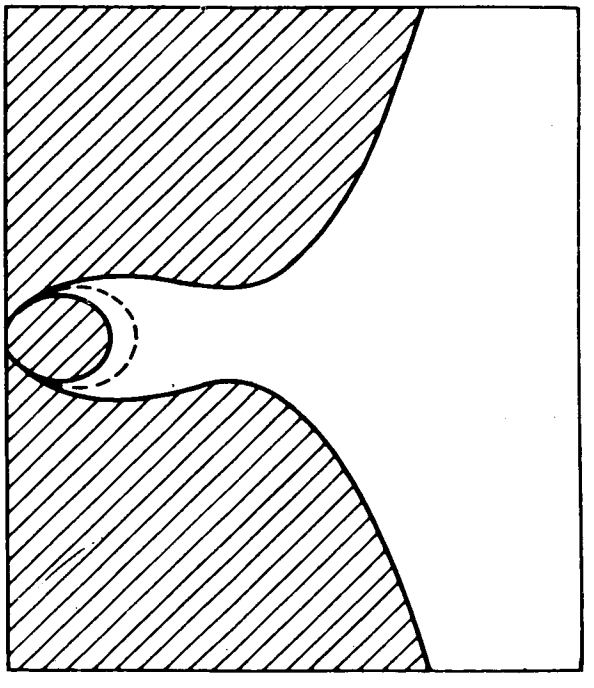


FIGURE 1.—A Størmer diagram.

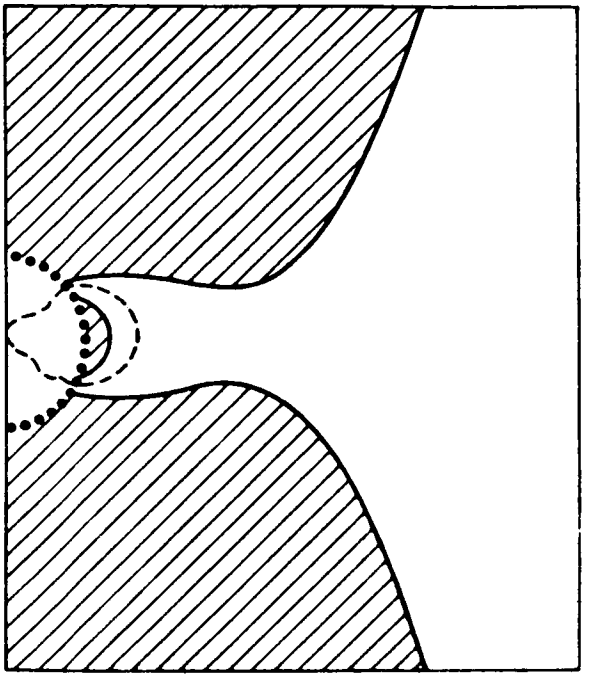


FIGURE 2.—A Størmer diagram when the field near the earth is not axially symmetric.

where r is in units of $M^{1/2}R^{-1/2}$ (Størmer units), M being the dipole moment, R the particle rigidity, and ϑ the colatitude. In the full three-dimensional configuration space, (2) is a surface with axially symmetry composed of lines of force. Since it is the only such surface which is entirely allowed, it is obvious that its lines of force are those about which the particle spirals, if one adopts the Alfvén description.

Figure 2 is a schematic representation of a different situation. Here, we suppose that outside the dotted circle the field is nearly that of a dipole. We then have allowed and forbidden regions as before. Inside it, we can draw only the line of force which, outside the dotted circle, is given by (2).

Let there be a particle spiraling about this line of force, and let it be such that when it moves in the region outside the dotted circle, $\gamma = 1$. This is just the critical value, since no particles with $\gamma > 1$ can reach the earth, by the well known analysis by Størmer. When $\gamma = 1$, (2) tells us that the line of force about which the particle spirals crosses the equator at $r = \frac{1}{2}$, in Størmer units. On the other hand, the point at which this line of force crosses

the equator is a distance L earth radii from the origin, where L is McIlwain's parameter.[5] This is so because the field in that region is that of a dipole. Consequently, for the particle in question we know that a particular distance is $\frac{1}{2}$ Størmer unit and also L earth radii. Upon multiplying $\frac{1}{2}$ by the conversion factor to earth radii, equating the result to L, and squaring both sides, we obtain (1). It is obvious from the arguments leading to this result that L must be calculated at the guiding center rather than a point on the trajectory.

A completely analogous argument can be made whenever the field at suitably large values of r is axially symmetric, even when it is not that of a dipole. Instead of $\frac{1}{2}$ and L, one will then obtain two other expressions for the distance in question which lead in the same way to an analogue of (1).

One can perhaps doubt that the particle really spirals about the line of force given by (2). The rest of this section will demonstrate that result.

In the equatorial plane at distances large enough that the field is that of a dipole, we have, from Størmer's integral[6]

$$\cos\omega = 2\gamma r^{-1} - r^{-2}.$$

On solving this for γ, we have

$$2\gamma = r^{-1}(1 + r^2\cos\omega). \tag{3}$$

Let a be the larmor radius in Størmer units and α the pitch angle of the spiral trajectory. It is then a simple algebraic result that

$$a/r = r^2\sin\alpha. \tag{4}$$

Let ψ be azimuth about the guiding center line of force. Then at the equator, making an error of order a/r,

$$\cos\omega = \sin\alpha\cos\psi. \tag{5}$$

Also, we have exactly

$$r = r_0 + a\cos\psi \tag{6}$$

where r_0 is the radial distance to the guiding center. From (4) and (6), we have

$$r = r_0 + r^3\sin\alpha\cos\psi. \tag{7}$$

From (3) and (5),

$$2\gamma = r^{-1}(1 + r^2\sin\alpha\cos\psi). \tag{8}$$

Now,

$$r_0/r = 1 - (r - r_0)/r.$$

From (7) this becomes

$$r_0/r = 1 - r^2\sin\alpha\cos\psi.$$

Then (8) becomes

$$2\gamma = r_0^{-1}[1 - (r^2\sin\alpha\cos\psi)^2]. \tag{9}$$

From (4) it is evident that, dropping a term of order $(a/r)^2$, this becomes

$$2\gamma = r_0^{-1}. \tag{10}$$

Since the critical value of γ is unity, (10) shows that $r_0 = \frac{1}{2}$, the desired result.

A simple numerical estimate for the case of a dipole magnetic field shows that if the geomagnetic latitude is greater than 45°, the neglected term in (9) is less than 0.008. This result is devoid of any great interest, since the trajectory is not described to second order in (a/r).

COMPARISON WITH PREVIOUS RESULT

There are three sources of the discrepancy between (1) and the cutoffs computed inreference 1 by the machine method. (See figure 6 of ref. 1.) In the first place, L was previously computed at the observation point rather than the guiding center since the significance of (1) had not been learned at that time. This procedure introduces important errors only at the lower latitudes. In the second place, the previous machine calculations contain an error which is significant at all latitudes. This error arose from applying the Størmer criterion at too small a radius. While the nondipole terms in the expansion contribute one per cent or less to the total magnetic field at the place in question, they have a considerably more significant effect than this on the particle motion. Table I exhibits the effect of this error. The first two columns of rigidity values give the values obtained from Eq. (1) with L computed according to the following scheme. The line of force was computed numerically. For the first column it was run to two earth radii from the center of the earth, and L was set equal to $2/\sin^2\theta$, where θ is the

TABLE I.—*Rigidity Cutoffs at* 70°N *Geographic Latitude* (Bv)

East longitude (deg)	Two earth radii	Min. of B	Machine
0	0.210	0.261	0.212
40	0.453	0.454	0.456
80	0.628	0.526	0.630
120	0.708	0.581	0.712
160	0.635	0.570	0.639
200	0.329	0.318	0.330
240	0.062	0.060	0.061
280	0.006	0.007	0.006
320	0.040	0.058	0.039

geomagnetic colatitude. For the second column the line was run to the point where B passes through its minimum value, and L was set equal to r at that point. In the third column of the table appear cutoffs obtained by the previous machine method and with the Størmer criterion applied at two earth radii.

We claim that the first and third columns of Table I, by their near identity, show that (1) and reference 1 produce the same cutoffs when both are used in the same (somewhat incorrect) way.

There is a third source of discrepancy between values obtained by the two methods as reported in reference 1. The machine calculated cutoffs were obtained for the Finch and Leaton field, while L was computed for the Jensen and Whitaker field.[7] In Table II we give L values computed by numerically integrating along the field line to the point where the magnetic field has its minimum and setting L equal to the distance from the origin to that point. The three columns respectively give L for the Jensen and Cain field, the Finch and Leaton field, and the Jensen and Whitaker field.[8] It should be noted that Jensen and Whitaker included a monopole term in their expansion, which term cannot occur in nature. The inclusion of this term of course strongly changes all the high order terms from what their values would be in its absence.

We can obtain some estimate of the error arising from this source in the following way. In a magnetic field consisting of a monopole plus dipole only, the scalar potential is

$$V=(Q/r)+(M\cos\vartheta/r^2).$$

By solving the differential equations of a line of force, we obtain

$$L^{-1}=(\sin^2\vartheta/r)+(Q/M)\cos\vartheta.$$

The Jensen-Whitaker field has $Q/M\cong 6\times 10^{-4}$. Let $\vartheta=15°$, $r=1$ earth radius. Then $L=14.8$. If the monopole term is omitted, $L=14.9$. The difference is less than 1%. Thus, while the presence of the monopole of course vitiates all of the high order terms in the expansion, the error introduced in our case is not important. Evidently the differences in Table II arise from changes in the location of the roots of lines of force rather than in changes of their large scale structure.

TABLE II.—*L Values at* 70°N *Geographic Latitude*

East longitude (deg)	L_1 Jensen Cain	L_2 Finch Leaton	L_3 Jensen Whitaker	$(L_2-L_1)/L_2$ (%)	$(L_2-L_3)/L_2$ (%)
0	7.59	7.55	7.29	−0.53	3.5
40	5.85	5.73	5.62	−2.1	1.9
80	5.43	5.32	5.35	−2.1	−0.6
120	5.11	5.06	5.24	−1.0	−3.6
160	5.15	5.11	5.30	−0.8	−3.7
200	6.95	6.85	7.09	−1.5	−3.5
240	16.10	15.80	16.00	−1.9	−1.3
280	43.5	45.0	39.9	3.3	11.3
320	15.78	16.01	14.98	1.5	6.4

TABLE III.—*Machine Computed Cutoffs*

Location of field line at earth's surface		Cutoff rigidity (Bv) in the magnetic field direction at geocentric distances of—						*L*-derived cutoff (Bv) R_L
Long. East (deg)	Latitude (deg)	$r=r_e$	$1.5r_e$	$2.0r_e$	$2.5r_e$	$3.0r_e$	$3.5r_e$	
100	66.0	0.934	0.934	0.934	0.933	0.933	0.924	0.916
	60.7			1.63	1.63	1.63		1.60
	51.5	3.70		3.70				3.82
	−12.7	12.40						12.32
	−25.7	6.32						6.56
	−33.7	3.62	3.63	3.63				3.76
	−43.0	1.40	1.40		1.40	1.43		1.39
	−47.5	0.812	0.813	0.814	0.815	0.816	0.817	0.819
200	62.5	0.940	0.942	0.942	0.943	0.943	0.945	0.943
	57.5	1.65	1.66		1.66	1.63		1.60
	47.5	3.83		3.89				3.75
	19.3		12.25					12.43
	−35.0		6.70					6.56
	−44.8	3.54						3.62
	−54.2	1.66	1.67		1.65	1.64		1.62
	−60.1	0.915	0.916	0.916	0.916	0.915	0.911	0.895
300	49.0	0.900	0.897	0.895	0.895	0.893	0.896	0.885
	42.7	1.72	1.71	1.70	1.68			1.69
	32.2	3.60	3.70	3.58				3.68
	−70.8				1.72			1.79
	−75.9	1.01	1.01	1.00	1.00	1.00	1.01	1.03

LOW LATITUDE CUTOFFS

Table III contains values of the cutoff rigidity R_B in the direction of the local Finch and Leaton field, calculated by numerically integrating the equations of motion. The results at low latitude ($R>3$ Bv) are sparse because of the length and complexity of the particle trajectories involved. It will be seen, however, that where several values of R_B along a field line have been obtained, these values are indeed approximately invariant. In addition, all of the values of R_B computed are consistent with (1).

REFERENCES

1. SAUER, H. H., *J. Geophys. Research*, **68,** 957 (1963).
2. ALFVEN, H., "Cosmical Electrodynamics," sections 2.3. Oxford Univ. Press, 1950.
3. FINCH, H. F., and LEATON, B. R., *Monthly Notices Roy. Astron. Soc., Geophys. Suppl.* **7,** 314 (1957).
4. STØRMER, C., "The Polar Aurora," fig. 123. Oxford Univ. Press, 1955.
5. MCILWAIN, C. E., *J. Geophys. Research*, **66,** 3681 (1961).
6. STØRMER, C., "The Polar Aurora," p. 218. Oxford Univ. Press, 1955.
7. JENSEN, D. C., and WHITAKER, W. A., *J. Geophys. Research*, **65,** 2500 (1960).
8. ———, and CAIN, J. C., *J. Geophys. Research*, **67,** 3568 (1962).

THE LOW FREQUENCY POWER SPECTRUM OF COSMIC-RAY VARIATIONS DURING IGY

DAVID STERN*

Goddard Space Flight Center

The power spectra of cosmic-ray variations, as recorded by neutron monitors at various locations, is computed and analyzed herein. The frequency range of the variations covered (from 0.005 to 0.5 per day) is found to contain only one significant peak, contributed by the well known 27-day variation. By use of the power spectrum, a quantitative estimate of the latitude dependence of the 27-day variation may be made. From this, the average rigidity dependence of the modulation amplitude is deduced, and is found to be essentially the same as that of Forbush decreases. The average was taken over the period July 1957 to December 1958, during which the peak decreased significantly.

Irregular variations were also investigated, as was the correlation with magnetic activity. It is found that the irregular variations have approximately the same latitude dependence as the 27-day peak, decreasing roughly exponentially with increasing frequency. There was some correlation between magnetic activity and cosmic-ray variations, but it showed no clear recurrence tendencies.

A general review of power spectral analysis, with emphasis on points relevant to this work, is included here, and the conclusions along with the connection between Forbush decreases and the 27-day variation are discussed.

INTRODUCTION

Examination of the record of a high counting rate cosmic-ray neutron monitor immediately reveals its large variability. At least four definite sources of variation can be traced: the 11-year cycle; the Forbush decreases accompanying sudden commencement (SC) magnetic storms; the 27-day variation; and the daily variation. Of these, three are periodic in character: (1) the 11-year variation, which is connected with the cycle of solar activity and will not be discussed further; (2) the 27-day variation, which is believed to be associated with the period of solar rotation; and (3) the daily variation, which reflects anisotropy in the cosmic radiation reaching the earth. Both the 27-day and the daily variation are often observed to undergo large changes in amplitude within a few cycles. In addition, the 27-day variation seems to vary in both phase and frequency.[1–3] If the sources of the 27-day periodicity are located upon the sun, such changes are indeed expected since all features of the solar surface have a transient nature. Thus, a source of 27-day variation may in time be superseded by another, at a different latitude (which affects the rotation period), and at a different longitude (which affects the phase). Changes of this kind, unfortunately, complicate the study of the phenomenon.

Apart from these well-defined variations, the counting rate undergoes various irregular fluctuations. Comparison between stations confirms the fact that these are genuine fluctuations, and not of instrumental origin. Some of the questions associated with these fluctuations are:

(1) What is their typical time scale?
(2) What is their energy dependence?
(3) Do there exist in the primary flux any periodic variations other than those listed?

To answer these and similar questions, and to

*This work was performed under a National Academy of Sciences postdoctoral scholarship.

analyze phase-unstable periodic variations quantitatively, the method of power spectrum analysis is very useful. This method will now be briefly reviewed.

REVIEW OF POWER-SPECTRAL ANALYSIS

Conventional Fourier analysis is not suitable for the analysis of a time-dependent counting rate $x(t)$, which tends to be periodic with frequency f_0, but randomly changes its phase, now and again. Indeed, it can be shown that as the length of the given record tends to infinity, the corresponding estimated Fourier transform of $x(t)$ for any frequency approaches zero. This is true even for f_0; over the long run, $x(t)$ will have equal probability for being in or out of phase. Two main approaches exist for frequency analysis in this case.

One approach is based on the fact that even though the Fourier transform of $x(t)$ approaches zero in the limit, its mean square, under very general assumptions, tends to a finite limit $P(f)$:

$$P(f) = \lim_{T\to\infty} \frac{1}{T} \left| \int_{-T/2}^{T/2} x(t) e^{2\pi f i t}\, dt \right|^2 .$$

The function $P(f)$ is called the power spectral density, and will be strongly peaked at $f=f_0$. It differs only in minor details from the "Periodogram" introduced by A. Schuster (reference 4; also see, for instance, reference 5, Section 16.30). It can be intuitively understood as follows: let $x(t)$ be regarded as a voltage signal, and let it be passed through a filter network sharply tuned to pass only a narrow band Δf around the frequency f. Let the output signal be fed into a resistance of 1 ohm. The mean power of the output signal, which is proportional to the mean square of the output voltage, will then be $P(f)\,\Delta f$ (hence the name *power spectrum*). Power spectral analysis of a time-varying quantity is, therefore, similar to analysis of an unknown voltage signal for its frequency content by means of frequency filters.[6]

An alternative approach is the investigation of recurrence tendencies. An early method attributed to Chree[7] consists of selecting, according to a predetermined criterion, times at which $x(t)$ was highest (or lowest), and then superimposing the record upon itself so that all selected points overlap. If there is a tendency for a maximum (or minimum) of $x(t)$ to recur after a period τ, this will generally show up in the sum of the superimposed records. This method has clearly demonstrated, among other things, the 27-day periodicity in cosmic-ray variations (see, for example, references 2, 3, 8, 9, 10, and 11). Unfortunately, it is not suitable for quantitative evaluations.

A more satisfactory measure for recurrence is the autocorrelation or autocovariance function, defined as:

$$C(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} [x(t) - \bar{x}]\,[x(t-\tau) - \bar{x}]\, dt .$$

If we first normalize $x(t)$ so that $\bar{x}=0$, then $C(\tau)$ simplifies to

$$C(\tau) = \overline{x(t)\, x(t-\tau)} .$$

Assuming that the process is symmetric about its mean, $x(t)$ has equal probability to be of either sign. The same holds for $x(t-\tau)$, provided it is totally unrelated to $x(t)$; in this case $C(\tau)$ will, over a long run, approach zero. On the other hand, if τ represents a recurrence period, whatever the sign of $x(t)$, then $x(t-\tau)$ will have more than even probability of being of the same sign, so that $C(\tau)$ tends to a positive limit.

Wiener and Khintchine (reference 12, footnote 16) showed that the two measures described here contain equivalent information, and each can be derived from the other by means of a Fourier transformation:

$$C(\tau) = \int_{-\infty}^{\infty} P(f) e^{2\pi i f \tau}\, df \;;$$

$$P(f) = \int_{-\infty}^{\infty} C(\tau) e^{-2\pi i f \tau}\, d\tau .$$

The ordinary method of estimating the power

spectrum, though not the only one (see, for instance, reference 13), therefore involves prior estimation of the autocorrelation function by means of the finite record on hand, from which an estimate of $P(f)$ is obtained by transformation. Though both functions contain equivalent information, it is useful to consider both for complete understanding of the behavior of $x(t)$. If there is a recurrence tendency with period τ_0 lasting more than one cycle, $C(\tau)$ will have peaks not only at τ_0 but also at $2\tau_0$, $3\tau_0$, etc.; it will, therefore, indicate the average number of oscillations between phase jumps. On the other hand, $P(f)$ has the advantage of concentrating all the information about the component with frequency f at one point. This is especially important when there is more than one frequency involved, in which case, $C(\tau)$ is often rather irregular.

More details can be found in a number of books and reviews dealing with the subject.[12,14–20] Of these, the one by Blackman and Tukey[18] is most useful in dealing with the practical problems connected with actual estimation of the power spectrum; the present computation, using SHARE program 574 adapted for the 7090 computer, essentially follows their method. Some relevant points, quoted here without details, are the following:

Frequency Resolution.—If the data are sampled at intervals Δt, $C(\tau)$ can only be estimated for integral multiples of Δt. Furthermore, if $\tau_{max} = m\,\Delta t$ so that $C(\tau)$ is estimated for $m+1$ values of τ, the resulting estimates of $P(f)$ will cover $m+1$ points, equally spaced in frequency, from zero to $f_{max} = \frac{1}{2}\Delta t$. Each estimate of $P(f)$ will represent an average of the power spectrum over a band of the order $f_{min} = \frac{1}{2}m\,\Delta t$ around the frequency it represents; the exact width and shape of the band depend on the relative weights given to the estimates of $C(\tau)$. "Hanning windows" were used in this case. In the present computation, daily averages of the cosmic-ray intensity were used and usually 100 estimates were taken, covering the spectrum for periods between 2 and 200 days.

Aliasing.—If the power spectrum does not vanish above f_{max}, higher frequencies will contribute to the estimated spectrum in a way which cannot be resolved (aliasing). The spectrum of cosmic-ray variations has been found to fall off with increasing frequency rapidly enough as to make this source of error negligible. Even if this were not so, the fact that data points represent daily *averages*, and not momentary samples taken at daily intervals, strongly suppresses contributions from higher frequencies.

The Accuracy of Estimation.—The method here described assumes that the process is stationary in time—i.e., insensitive to a shift of time axis, and this assumption may be only an approximation. Indeed, the power spectral density of cosmic-ray variations changes over the period investigated, and therefore, any estimate of it gives only a time-averaged result.[21] Secondly, the question arises as to how closely the spectral estimates drawn from a limited sample approximate the actual values. Blackman and Tukey show that the ratio between a sample obtained by m estimates from n data points, and the actual value, approximately follows a χ^2 distribution with $\nu = 2n/m$ degrees of freedom. For analyses covering the IGY period (July 1957 to December 1958), $n = 549$, giving $\nu = 11$. Finally, it should be remembered that because of the statistical fluctuations in the counting rate, our record is not limited in length, but also has not been sampled with ideal accuracy. It can be shown, however, (Appendix A) that at least in the present case, this source of error is totally negligible.

Power spectrum analysis has been used in the investigation of diverse processes, such as the free oscillations of the earth, recurrence of magnetic storms, frequency analysis of sea waves, turbulence, and many others. Only in a few cases, however, has it been used to analyze cosmic-ray variations. Fonger[22] used the autocorrelation to demonstrate a 27-day recurrence, but since his data covered only a three month period, they are not very significant statistically. Panofsky, Lethbridge and Neuberger[23] obtained power spectra of neutron monitor rates, and cross correlated them with various meteorological data. In the present work, an attempt has been made to deduce detailed properties of low frequency variations of the cosmic-ray intensity (as measured by neutron monitors during IGY), and especially of the 27-day variation.

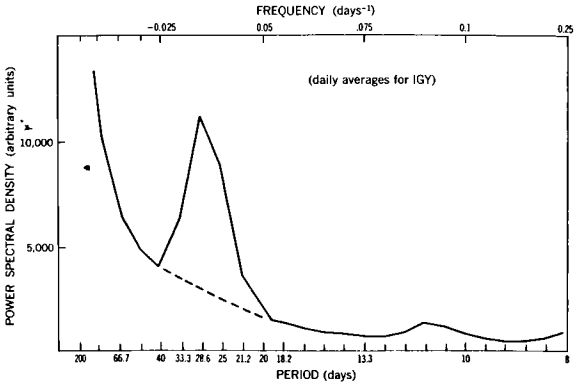


FIGURE 1.—The low frequency power spectrum of cosmic-ray variations during IGY as recorded by the Climax neutron monitor.

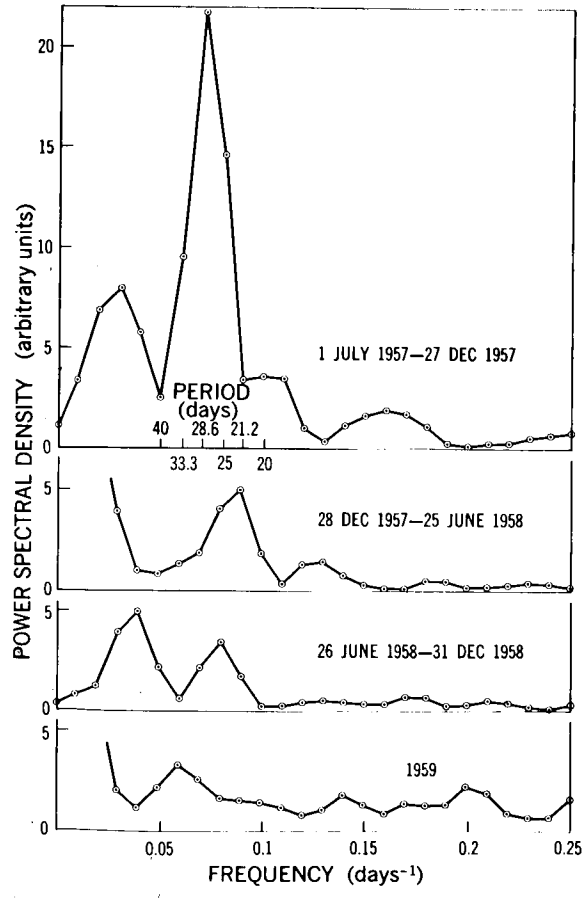


FIGURE 2.—The power spectrum of cosmic-ray variations as recorded by the Zugspitze neutron monitor for different parts of IGY and for 1959.

THE POWER SPECTRUM AT LOW FREQUENCIES

As figure 1 shows, the power-spectral density exhibits a marked peak between 27 and 28 days. The location of the peak is not fixed (see figure 2), and over the period July 1957–December 1959 its variation does not exhibit any marked trend (cf. reference 2, table III). None of the other peaks in the spectrum is believed to be significant; in particular, the absence of a conspicuous second harmonic is noted. Throughout the period

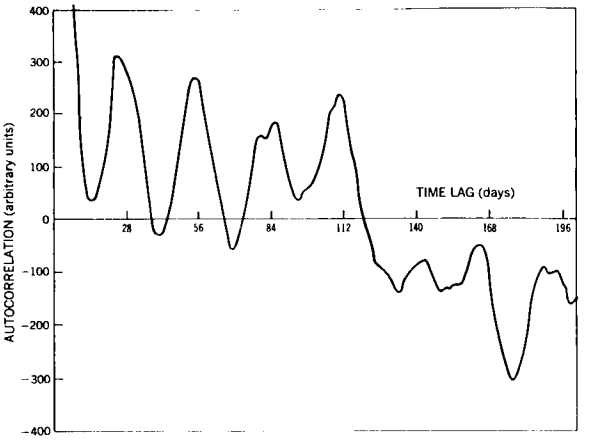


FIGURE 3.—The autocorrelation function of cosmic-ray variations during IGY as recorded by the Climax neutron monitor.

analyzed, the amplitude of the 27-day peak rapidly decreases with time, as can be seen in figure 2. The results described here are averages computed over the IGY period; during the first half year of the IGY, the amplitudes were about twice this average magnitude.

The autocorrelation has been plotted at daily intervals for 200 days (see figure 3), and it is evident that the recurrence tendency is relatively stable.

The counting rate was also cross-correlated with the magnetic activity C indices, tabulated in the reports of Solar-Geophysical data by the National Bureau of Standards. A sharp negative peak is observed in the cross-correlation function:

$$C_c(\tau) = \overline{X(t)\,Y(t-\tau)}\,,$$

where $X(t)$ is the counting rate and $Y(t)$ is the magnetic activity index. It is located near $t=1$

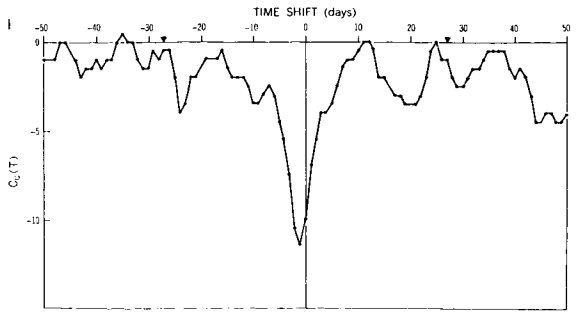


FIGURE 4.—The cross correlation function between magnetic activity and cosmic-ray intensity. On the right-hand side, cosmic-ray intensity precedes magnetic activity; on the left, the order is reversed.

day (see figure 4), indicating that high geomagnetic activity is likely to be followed within one day (on the average) by low cosmic-ray rates. This agrees with known properties of SC magnetic storms; a 27-day recurrence tendency was not conspicuous in this case. There seems to be, on the whole, quite a difference between the 27-day recurrence tendencies of cosmic radiation and that of magnetic activity.[23,24] While the spectrum of cosmic-ray variations shows a single well defined peak, that of magnetic activity not only has a second harmonic often exceeding the fundamental, but also shows higher harmonics, up to the sixth (figure 5).

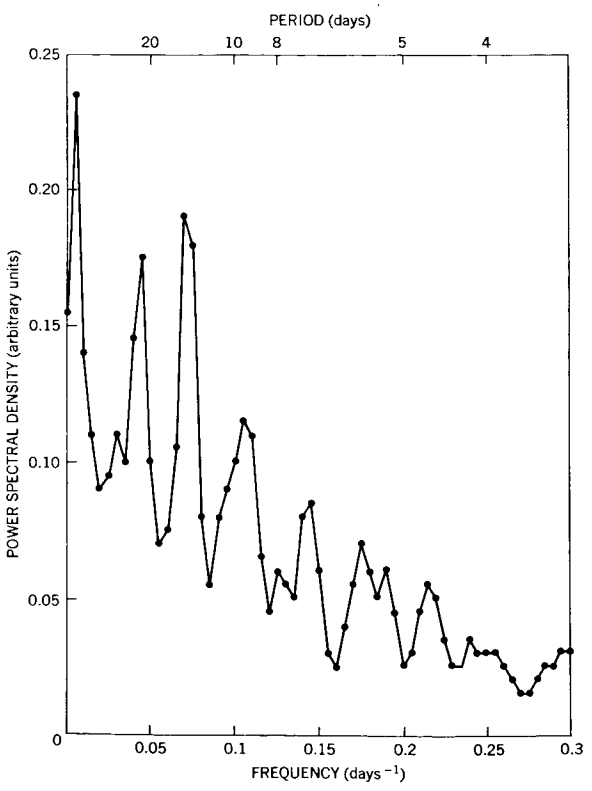


FIGURE 5.—The power spectrum of magnetic activity indices during IGY.

THE RIGIDITY DEPENDENCE OF THE 27-DAY VARIATION

Regarding the counting rate as a voltage signal, we can define the "power" contained in it, between the frequencies f_1 and f_2 as the area contained by the power-spectral density graph between the two frequencies:

$$\text{Power} = \int_{f_1}^{f_2} P(f)\, df .$$

The power contained in the 27-day variation will be proportional to the area enclosed by the 27-day peak. It should be noted however, that this quantity will not be obtained by passing the data through a filter which selects a limited band around 27 days. In that case, the power would also

TABLE 1.—*Average Amplitudes of the 27-Day Variation for a Number of IGY Neutron Monitors*

Station	Amplitude (%)	Corrected amplitude (%)	Cutoff rigidity (Bev/c)
Mawson	1.59	1.59	0.2
Sulphur Mtn.	2.05	1.66	1.1
Uppsala	1.58	1.58	1.3
Mt. Wellington	1.62	1.50	2.0
Climax	1.84	1.305	3.0
Zugspitze	1.71	1.26	3.3
Rome	1.20	1.20	4.7
Hermanus	1.21	1.21	7.0
Alma Ata	0.94	0.85	7.3
Mt. Norikura	0.88	0.66	10.3
Huancayo	0.73	0.52	13.2
Lae	0.53	0.53	13.5

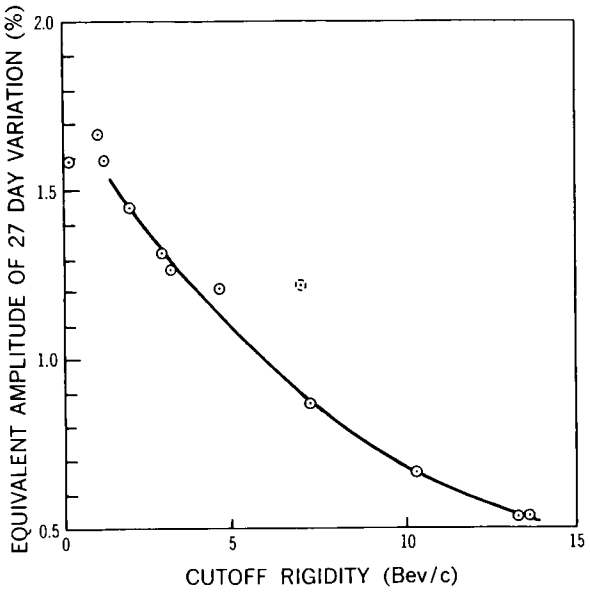


FIGURE 6.—Latitude dependence of 27-day amplitude, averaged over IGY.

contain a major contribution from the "pedestal" upon which the peak is superimposed, which presumably is due to irregular variations. We now define as the equivalent amplitude A of the variation the amplitude of a pure sinusoidal variation; containing the same power as is contained in the peak: the method by which A is calculated is described in Appendix B. The values of A thus obtained are between 0.5 and 2 percent.

The equivalent amplitudes have still to be corrected for altitude. Since the latitude dependdence of the variation strongly resembles that of Forbush decreases, we adopt the correction proposed for the latter case by McCracken and Johns,[25] amounting to 12 percent per 1000 meters. This is further justified by the fact that corrected amplitudes of some high altitude stations (e.g., Sulphur Mountain; Huancayo) fall close to those of sea level stations (e.g., Uppsala; Lae), having approximately the same cutoff. In all, 12 amplitudes were found and plotted against cutoff rigidities obtained from the eccentric dipole model by Kodama, Kondo, and Wada[26]; they are given in table 1 and in figure 6.

It will be seen that most of the points fall on a smooth curve, with the exception of Hermanus. This station is likely to have its cutoff lowered by local anomalies in the geomagnetic field, and was therefore not considered. It will also be noted that there seems to be no flattening of the curve down to a cutoff rigidity of about 1 Bev/c—this implies that the primary radiation in the low energy region undergoes very large variations.

From the latitude dependence of the counting rate, it is possible to deduce the modulation experienced by various portions of the primary spectrum. Let

$N(p)$ = the sea level counting rate at cutoff rigidity p,
$\delta N(p)$ = the 27-day equivalent amplitude,
$S(p)$ = the primary differential proton spectrum,
$\delta S(p)$ = the amplitude of the 27-day variation undergone by the primary spectrum at rigidity p, and
$Y(p)$ = the "gross" yield function at rigidity p [reference 27, equation (11)].

The experimental data consist of the equivalent amplitudes $A(p)=\delta N/N$, and the latitude dependence $N(p)$. Neglecting penumbral effects, we can write

$$N(p) = \int_p^\infty S(p')\,Y(p')\,dp', \qquad (1)$$

$$\delta N(p) = \int_p^\infty \delta S(p')\,Y(p')\,dp' = N(p)\,A(p) \qquad (2)$$

Differentiating these expressions, and dividing equation (2) by equation (1), we have

$$\frac{\delta S}{S} = \frac{d}{dN}\,[N(p)\,A(p)]$$

We now need two experimentally determined functions of the cutoff rigidity—the equivalent amplitude $A(p)$, and the total counting rate $N(p)$. Unfortunately, not many sea level measurements of the latitude dependence of $N(p)$ exist for IGY. We shall, therefore, base our calculations on the quiet-time rates $N_0(p)$ used by Quenby and Webber[27]; they are arbitrarily normalized to $N_0(15\ \text{Bev})=100$.

Let us denote the ratio between the differential primary spectra during IGY and during the solar

TABLE 2.—*The Rigidity Dependence of the 27-Day Variations*

p (Bev/c)	N_0	$\frac{dN_0}{dp}$	α	$10^2A(p)$	$\frac{dA}{dp}$	$\delta S/S$ uncorrected (%)	$\delta S/S$ corrected (%)
2.5	172.5	3.2	0.58	1.38	0.135	8.79	14.0
3.5	168.0	5.8	0.68	1.26	0.110	4.44	5.93
4.5	161.3	7.6	0.76	1.15	0.115	3.57	4.38
5.5	153.4	8.2	0.81	1.045	0.105	3.00	3.48
6.5	145.2	8.1	0.84	0.935	0.100	2.73	3.08
7.5	137.5	7.5	0.855	0.843	0.080	2.32	2.56
8.5	130.3	6.7	0.87	0.770	0.070	2.13	2.34
9.5	124.0	6.0	0.875	0.705	0.060	1.93	2.12
10.5	118.4	5.2	0.88	0.648	0.055	1.90	1.98
11.5	113.5	4.6	0.89	0.603	0.045	1.72	1.85
12.5	109.1	4.1	0.895	0.555	0.040	1.61	1.74

minimum (1954–5) by

$$\alpha(p) = \frac{S(p)}{S_0(p)} ;$$

values of $\alpha(p)$ are taken from the work of F. McDonald.[28] The total counting rate $N(p)$ changes relatively little over the solar cycle; therefore, no correction is applied to it. On the other hand, dN/dp is proportional to $S(p)$, and receives a correction factor α. Substituting in equation (3), we obtain

$$\frac{\delta S}{S} = A(p) + N_0(p)\frac{\frac{dA}{dp}}{\alpha(p)\frac{dN_0}{dp}} .$$

The results are presented in table 2 and in figure 7.

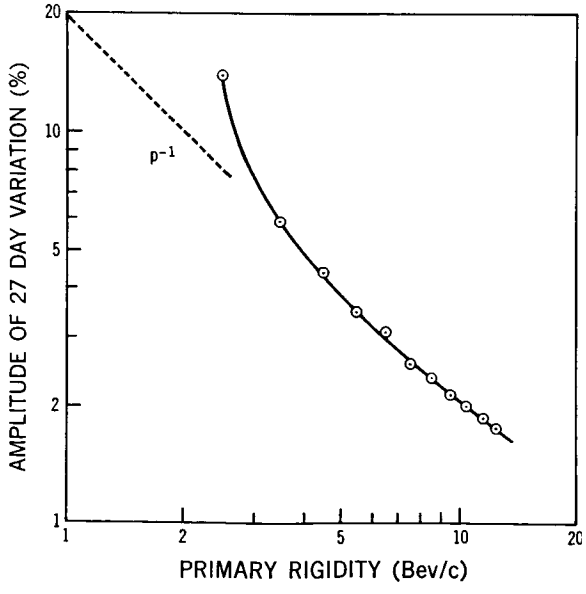


FIGURE 7.—Rigidity dependence of the 27-day modulation undergone by the primary spectrum, averaged over IGY.

The rigidity dependence obtained here may be compared to other experimental results. It can be seen from figure 7 that between 12.5 and 3.5 Bev/c, it roughly follows a p^{-1} relation; this agrees with results found for Forbush events.[29] The dependence does not seem to be a power-law one, however, and it steepens at low rigidities. The rigidity dependence should also be compared with the large 27-day amplitudes found by other workers, especially by the Russians[30–33] at moderate balloon altitudes and high latitudes. Since balloon-borne instruments are relatively more sensitive to low rigidities, their results are in general agreement with the results obtained here. We cannot very well extrapolate the rigidity dependence to low energies; it seems, however, quite possible that the low energy cutoff of primary cosmic radiation undergoes a 27-day modulation at solar maximum, and that this may be a cause for experimental discrepancies.

IRREGULAR VARIATIONS

The 27-day peak is superimposed upon a continuum, representing the contribution of irregular

TABLE 3.—*Latitude Dependence of Irregular Variations in Cosmic Ray Intensity During IGY*

Station	Power spectral density/area of 27-day peak (arbitrary units)						
	Freq. = 0.005/day	0.010/day	0.015/day	0.020/day	0.055/day	0.060/day	0.065/day
Mawson	1095	358	272	277	106	71	48
Sulphur Mountain	930	477	305	219	81	47	37
Uppsala	734	409	341	292	91	70	55
Mt. Wellington	958	440	327	282	107	65	48
Climax	919	467	358	316	93	66	53
Zugspitze	753	351	314	247	107	73	49
Rome	989	425	281	232	109	75	52
Hermanus	5635	1357	422	333	215	133	51
Alma Ata	9042	1886	1108	810	316	381	89
Norikura	1495	513	363	334	177	98	70
Huancayo	1808	422	340	362	158	66	53
Lae	2777	1309	679	696	198	214	175

variations to the power spectrum. This continuum (plotted against frequency) is roughly exponential in shape, reaching a relatively constant noise level (see Appendix B) at a frequency of about 5/day.

The latitude dependence of the continuum seems to be approximately the same, at first glance, as that of the 27-day peak. This in itself is not surprising, since it is bound to contain a major contribution from Forbush events, which have practically the same latitude dependence. The comparison is complicated, however, by the fact that the continuum also contains contributions from instrumental drifts, which are not latitude dependent; this will tend to make the latitude dependence less steep than that of the 27-day variation. Because of this effect, one should discount stations in which the continuum is abnormally high.

Table 3 gives the ratio (in arbitrary units) of the power-spectral density to the area of the 27-day peak for various frequencies. If the continuum has the same latitude dependence as the 27-day variation, this ratio at any given frequency should be constant for all stations. It can readily be seen that for three stations (Lae, Hermanus and Alma Ata) the ratios are unusually large; it is possible that these stations experience considerable drifts. For the other stations, there may be some increase towards the equator, but generally, the ratio seems to be fairly constant. It is, therefore, reasonable to assume that the latitude dependence of the continuum at low frequencies is close to that of the 27-day peak.

CONCLUSION

Several theories exist about the cause of the 27-day variation. One approach has been investigated in detail by Alfvn.[34,35] Alvn assumes that the interplanetary magnetic field near the solar equatorial plane contains "beams" of high plasma flux and magnetic field density, which co-rotate with the sun. These beams last for several rotations, and every time they intercept the earth, magnetic and cosmic-ray disturbances occur.

Alternative approaches are discussed by Dorman, who examined and rejected various explanations based on (1) a solar magnetic dipole noncoincident with the solar rotation axis, (2) atmospheric effects, and (3) high energy particles produced by solar flares. In reference 37, Section 32, paragraph 9(d), he suggests that "the effect of the decrease in cosmic-ray intensity during the

time of geomagnetic disturbances, is the basis of the phenomenon of 27-day variations of the cosmic rays." The Forbush decreases, to which Dorman refers, are obviously nonperiodic phenomena, as they can generally be traced back to solar flares occurring a day or two previously. However, flares tend to be associated with centers of solar activity, and these are not evenly distributed in solar longitude. As is shown by the 27-day variation of sunspot numbers, there will generally be one center much more active than the rest. Every time this center faces the earth, there is a marked tendency for sudden commencement-type magnetic storms and their associated Forbush decreases to occur, leading to an apparent 27-day periodicity.

The results obtained here support this hypothesis. The energy dependence of the variation approximates that obtained for the Forbush effect.[29,37] The variation in period length (see figure 2) can be explained by noting that the "favorable interval" for an active area to cause a Forbush decrease is quite wide[38]; the cosmic-ray record for the beginning of IGY (figure 8), when the 27-day variation was very high, shows that the separation of the main Forbush events was indeed of the order of 27 days.

Examining this point more closely, we find that the main Forbush events alone cannot account for the 27-day variation. For one thing, the second harmonic of the variation would have been higher in that case. We also note that the cosmic-ray intensity (figure 8) often starts decreasing before the main Forbush event. These "predecreases"[39] may be tentatively identified as Forbush decreases, the main impact of which misses the earth; they cause considerable smoothing of the 27-day variation, especially at high altitudes (reference 30, figure 1).

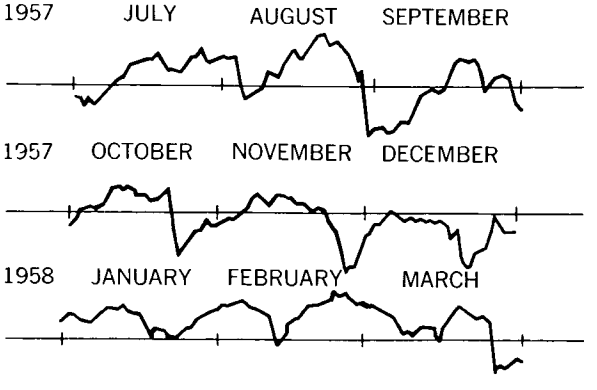


FIGURE 8.—Cosmic-ray intensities during the first half of IGY as recorded by the Climax neutron monitor.

When we take into account the rigidity dependence of the 27-day variation obtained here, its correlation with magnetic activity, and its time dependence, it appears very likely that this variation is associated with nonrecurrent magnetic storms, showing spurious periodicity because of their origin in solar activity centers. Perhaps recurrent magnetic storms may be responsible for a 27-day variation different from the one discussed here, with presumably a smaller amplitude (possibly conforming with Alfvèn's model). This can only be established by analyzing periods of low solar activity when Forbush events are rare and magnetic activity shows strong recurrence tendencies.[40]

REFERENCES

1. ZWANZIG, W., "Zur 27-tägigen Periode der Kosmischen Strahlung," *Zeitschrift fur Naturforschung*, **16**1(11): 1237–1239, November 1961.
2. VENKATESAN, D., "The 27-Day Variation in Cosmic Ray Intensity," *Supplemento al Nuovo Cimento*, **8**(2): 285–297, 1958.
3. ———, "Changes in Amplitude of the 27-day Variation in Cosmic Ray Intensity during the Solar Cycle of Activity," *Tellus*, **10**(1): 117–125, February 1958.
4. SCHUSTER, A., "The Periodicity of Sun Spots," *Astrophys. J.*, **23**(2): 101, March 1906.
5. CHAPMAN, S., and BARTELS, J., "Geomagnetism," Oxford: The Clarendon Press, 1940.
6. HOLLOWAY, J. L., Jr., "Smoothing and Filtering of Time Series and Space Fields," in: *Advances in Geophysics*, ed. by H. E. Landsberg and J. Van Mieghem, Vol. 4, pp. 351–389, New York: Academic Press, 1958.
7. CHREE, C., "Some Phenomena of Sunspots and of Terrestrial Magnetism at Kew Observatory," *Phil. Trans. Roy. Soc. London*, **A212**, 75–116, 1913.
8. MONK, A. T., and COMPTON, A. H., "Recurrence Phenomena in Cosmic-Ray Intensity," *Rev. Mod. Phys.*, **11**(3–4): 173–179, July–October 1939.
9. BROXON, J. W., "Relation of the Cosmic Radiation to Geomagnetic and Heliophysical Activities," *Phys. Rev.*, **62**(11–12): 508–522, December 1 and 15, 1942.
10. LOCKWOOD, J. A., and SHEA, M. A. (Univ. of New Hampshire, Physics Dept.), "The 27-Day Recurrent Variations in the Cosmic-Ray Intensity," Air Force Cambridge Research Laboratory-TN–60–699, 1960.
11. KODAMA, M., "The 27-Day Recurrence of Cosmic-Ray Intensity at the Minimum Solar Activity," in: *Proc. Moscow Cosmic Ray Conf.*, Vol. 4, pp. 287–292, Moscow, 1960.

12. RICE, S. O., "Mathematical Analysis of Random Noise," in: *Selected Papers on Noise and Stochastic Processes*, ed. by N. Wax, pp. 133–294, New York: Dover, 1954.
13. WELCH, P. D., "A Direct Digital Method of Power Spectrum Estimation," *IBM J. Res. and Dev.*, **5**(2): 141–156, April 1961.
14. BENDAT, J. S., "Principles and Applications of Random Noise Theory," New York: John Wiley and Sons, 1958.
15. DAVENPORT, W. B., and ROOT, W. L., "An Introduction to the Theory of Random Signals and Noise," New York: McGraw-Hill, 1958.
16. HANNAN, E. J., "Time Series Analysis," London: Methuen, 1960.
17. SOLODOVNIKOV, V. V., "Introduction to the Statistical Dynamics of Automatic Control Systems," Transl. ed. by J. B. Thomas and L. A. Zadeh, New York: Dover, 1960.
18. BLACKMAN, R. B., and TUKEY, J. W., "The Measurement of Power Spectra," New York: Dover, 1000.
19. VAN ISACKER, J., "Generalized Harmonic Analysis," in: *Advances in Geophys.*, ed. by H. E. Landsberg and J. Van Mieghem, Vol. 7, pp. 189–214, New York: Academic Press, 1961.
20. WIENER, N., "Extrapolation, Interpolation and Smoothing of Stationary Time Series," New York: John Wiley and Sons, 1949.
21. PAGE, C. H., "Instantaneous Power Spectra," *J. Appl. Phys.*, **23**(1): 103–106, January 1952.
22. FONGER, W. H., "Cosmic Radiation Intensity-Time Variations and Their Origin. II: Energy Dependence of 27-Day Variations," *Phys. Rev.*, **91**(2): 351–361, July 15, 1953.
23. PANOFSKY, H., LETHBRIDGE, M. D., and NEUBERGER, H., "Research Directed Towards the Study of the Relation of Solar Energy Variations to Changes of the Tropospheric Circulations," Penn. State Univ., Mineral Industries Exper. Sta., Final Rept. on Contract AF 19(604)–2251 (AFCRC–TR–58–263) June 30, 1958.
24. WARD, F. W., Jr., "The Variance (Power) Spectra of Ci, Kp, and Ap," *J. Geophys. Res.*, **65**(8): 2359–2373, August 1960.
25. MCCRACKEN, K. G., and JOHNS, D. H., "The Attenuation Length of the High Energy Nucleonic Component of the Cosmic Radiation near Sea Level," *Nuovo Cimento*, **13**(1): 96–107, July 1, 1959.
26. KODAMA, M., KONDO, I., and WADA, M., "Cut-off Rigidities of Cosmic-Ray Particles Calculated for the Eccentric Dipole Model of the Earth's Magnetic Field," *J. Scientific Research Institute* (Tokyo), **51**(1455–1464): 138–157, September 1957.
27. WEBBER, W. R., and QUENBY, J. J., "On the Derivation of Cosmic Ray Specific Yield Functions," *Phil. Mag.*, **4**(41): 654–664, May 1959.
28. MCDONALD, F. B., "Primary Cosmic-Ray Intensity near Solar Maximum," *Phys. Rev.*, **116**(2): 462–463, October 15, 1959.
29. LOCKWOOD, J. A., "An Investigation of the Forbush Decreases in the Cosmic Radiation," *J. Geophys. Res.*, **65**(12): 3859–3880, December 1960.
30. VERNOV, S. N., TULINOV, V. F., and CHARAKHCH'IAN, "The 27-Day Cosmic-Ray Period in the Stratosphere," *Doklady Akademii Nauk SSSR*, **122**(5): 788–791, October 11, 1958; Translation in *Soviet Phys.—Doklady*, **3**(5): 980–982, September–October 1958.
31. POMERANTZ, M. A., AGARWAL, S. P., and POTNIS, V. R., "Direct Observation of Periodic Variation of Primary Cosmic-Ray Intensity," *Phys. Rev.*, **109**(1): 224–225, January 1, 1958.
32. ———, ———, and ———, "Balloon Flight Investigations of Primary Cosmic Rays During Solar Disturbances," in: *Proc. Moscow Cosmic Ray Conf.*, Vol. 4, pp. 65–73, Moscow, 1960.
33. VERNOV, S. N., and SAMOSUDOV, B. E., et al., "Studies of Intensity Variations of Cosmic Radiation in the Stratosphere," in: *Proc. Moscow Cosmic Ray Conf.*, Vol. 4, pp. 53–64, Moscow, 1960.
34. ALFVÉN, H., "The Sun's General Magnetic Field," *Tellus*, **8**(1): 1–12, February 1956.
35. ———, "Interplanetary Magnetic Field," in: *Electromagnetic Phenomena in Cosmical Physics*, ed. by B. Lehnert, pp. 284–292, Cambridge: The University Press, 1958.
36. MCCRACKEN, K. G., "Energy Dependence of Transient Changes in the Primary Cosmic-Ray Spectrum," *Phys. Rev.*, **117**(6): 1570–1579, March 15, 1960.
37. DORMAN, L. I., "Cosmic Ray Variations," Moscow: State Publishing House for Technical and Theoretical Literature, 1957 (Translation by Technical Documents Liaison Office, Wright-Patterson Air Force Base, 1958).
38. SINNO, K., "Mechanism of Cosmic Ray Storms Inferred from Some Statistical Results," in: *Proc. Internat. Conf. on Cosmic Rays and the Earth Storm*, Kyoto, September 1961. II. Joint Sessions, pp. 395–399, Tokyo: Physical Society of Japan, 1962.
39. LEGRAND, J. P., "The Cosmic Rays 'Predecreases' (Prebaisse) in the Maximum Solar Activity Period (April 1957–December 1958)," in: *Proc. Moscow Cosmic Ray Conf.*, Vol. 4, pp. 219–227, Moscow, 1960.
40. TANDON, J. N., "Cosmic Ray Increases during the Sunspot Minimum of 1953," *J. Atmos. Terrest. Phys.*, **21**(2/3): 203–205, June 1961.
41. Japan, Science Council, National Committee for the International Geophysical Year, "Cosmic-Ray Intensity during the International Geophysical Year," Nos. 1–3, 1959–1960.

Appendix A

ESTIMATION OF THE FLUCTUATION NOISE

In practice, the cosmic-ray neutron rate $x(t)$ is estimated at fixed intervals Δt. Since the number of counts in any interval is subject to statistical fluctuation, this estimation is subject to error. We now wish to find the extent to which this will affect the estimation of power-spectral density $P(f)$.

Instead of dealing with an imperfectly sampled signal $x(t)$, it is found convenient to analyze a signal $x'(t)=x(t)+n(t)$ which is sampled without error—where $n(t)$ is the noise, duplicating the effects of statistical fluctuations. Since the measurements consist of averaging the counting

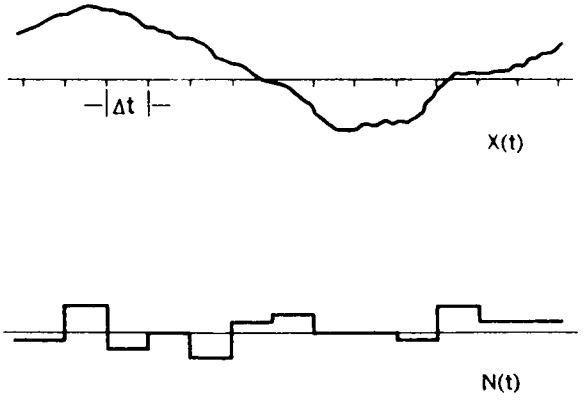


Figure B1

rate over time intervals Δt, we choose $n(t)$ to be a "histogram" function (see figure B1), each column of which has the width Δt. We assume that the heights of the columns are normally distributed around the mean; actually, they obey a Poissonian distribution, but it is only at very high sampling rates, when the average number of counts per interval is small, that the distinction is significant. Henceforth, we shall assume that both $x(t)$ and $n(t)$ have been normalized to average zero.

The noise is correlated with itself only for points within the same column; hence, its autocorrelation function is:

$$C_n(t) = \begin{cases} \overline{n_0^2}\,(1 - t/\Delta t) & \text{for } t < \Delta t. \\ 0 & \text{for } t > \Delta t. \end{cases}$$

The noise-power spectrum is then

$$P_n(f) = \int_{-\infty}^{\infty} e^{-2\pi i f t}\, C_n(t)\, dt,$$

$$= \overline{2n^2} \int_0^{\Delta t} \cos 2\pi f t \left(1 - \frac{t}{\Delta t}\right) dt,$$

$$= \overline{n^2}\, \Delta t\, [\mathrm{dif}\,(f\Delta t)]^2,$$

where

$$\mathrm{dif}\, x = \frac{\sin \pi x}{\pi x}.$$

is the well-known diffraction function.

Since the signal and noise are uncorrelated, the autocorrelation of $x'(t)$ will be the sum of those of the signal and of the noise, and the power spectrum will retain this additive property. The noise spectrum $P_n(f)$ is thus superimposed upon the desired spectrum. This in itself would cause no trouble if we had an infinite run of data at hand, since the true spectrum can then be obtained by subtracting the calculated value of $P_n(f)$. In a finite run, however, $P\nu(f)$ undergoes unpredictable fluctuations, as derived from a χ^2 distribution,[18] and it is this variability which constitutes the noisiness introduced by statistical fluctuations.

For the low frequency end of the spectrum, we may approximate $\mathrm{dif}(f\,\Delta t) \approx 1$. Let us assume that the cosmic-ray rate shows no variation apart from statistical fluctuations. Then, defining W, N_0,

and N, as in Appendix A, we have

$$\overline{n^2} = \overline{(W - \bar{W})^2}$$

and

$$\overline{N_1^2} = \left(\frac{N_0}{1000}\right)^2 (\overline{n^2}).$$

For purely statistical fluctuations, however,

$$\overline{N_1^2} = N_0,$$

hence,

$$\overline{n^2} = \frac{10^6}{N_0}.$$

The noise level will then be

$$P_n(f) \approx \frac{10^6 \Delta t}{N_0}.$$

The result is evidently inversely proportional to the counting rate. In actual practice, it seems to be somewhat larger than the above estimate indicates, probably because of nonstatistical fluctuations.

For daily averages of a typical neutron monitor station, such as Rome, $N_0 = 7.5 \times 10^5$. If time is measured in days, the noise level turns out to be of the order of unity. This is completely negligible in comparison to the amplitude of the 27-day peak, which on the same scale is close to 4500.

Appendix B

POWER SPECTRA OF LOGARITHMICALLY REDUCED DATA

The IGY data used in most of the present computation[41] do not give the counting rate N, but instead give

$$W(t) = 1000 \log\left(\frac{N}{N^*}\right) = 1000 \log N + \text{constant} ,$$

where N^* is a conveniently chosen constant. If we regard the counting rate as the sum of a constant rate N_0 and a small fluctuation N_1, averaging zero, we get (by expanding)

$$W \approx 1000 \frac{N_1}{N_0} + \text{constant} .$$

$$\overline{N_1^2} = \left(\frac{N_0}{1000}\right)^2 \overline{(W - \overline{W})^2} ,$$

$$= \left(\frac{N_0}{1000}\right)^2 C(0) ,$$

$$= \left(\frac{N_0}{1000}\right)^2 \int_{-\infty}^{\infty} P(f)\, df ,$$

$$= 2\left(\frac{N_0}{1000}\right)^2 \int_{0}^{\infty} P(f)\, df ,$$

where $C(t)$ and $P(f)$ are the autocorrelation and the spectral density obtained by analyzing $W(t)$. The power contained in any finite frequency band of $N(t)$ is proportional to the integral of P over the band, with the same proportionality factor as above.

Assume a frequency peak of area Q; if it were entirely due to a harmonic variation of frequency f and amplitude a, its power would equal $a^2/2$. Even if the peak is not infinitely sharp, we can define an equivalent amplitude a containing the same power; then

$$\left(\frac{a}{N_0}\right)^2 = 4 \times 10^{-6} Q .$$

In the text, relative equivalent amplitudes $A = a/N_0$ are usually given.

GEOMAGNETIC DISTURBANCES

MASAHISA SUGIURA*

Goddard Space Flight Center

With knowledge gained by satellite and space probe measurements substantial progress has been made in the understanding of geomagnetic disturbances. Much of the advancement in this field in the past three years probably results from our recognition of the importance of the magnetosphere in the interpretation of geomagnetic disturbances. In earlier years, energy sources of geomagnetic disturbances supplied by the Sun were thought to have their influences on the geomagnetic field almost directly without intermediate interactions in the magnetosphere.

The magnetic measurements made with instruments aboard Pioneer 1 [Sonnett et al., 1960], Pioneer 5 [Coleman et al., 1960], Explorer 10 [Heppner et al., 1963], and Explorer 12 [Cahill and Amazeen, 1963] indicate that steady solar wind causes the Earth's magnetic field to be confined in a cavity carved in the streaming solar plasma. At the subsolar point the interface between the magnetosphere and the solar plasma is situated at a geocentric distance of about 10 Earth radii. Within this interface the magnetic field is relatively quiet. In the region of thickness approximately 20,000 km outside the interface the magnetic field is irregular. Beyond this disturbed region the magnetic field is weak and less irregular.

According to the observations made by Explorer 10, the magnetospheric boundary on the dark side of the Earth appears to extend to 20 Earth radii or to even greater distances [Heppner et al., 1963].

*National Aeronautics and Space Administration—National Academy of Sciences—National Research Council Senior Research Associate on leave of absence from the Geophysical Institute, University of Alaska, College, Alaska.

The plasma measurement made with Mariner 2 by Neugebauer and Snyder [1962] established the existence of a steady flow of plasma from the Sun.

Coleman et al. [1960] observed an interplanetary magnetic storm by Pioneer 5 and related it to a magnetic storm observed on the Earth's surface.

Kellogg [1962] examined possible consequences of supersonic solar wind around the Earth and interpreted the outer boundary of the disturbed region outside the quiet magnetosphere as a shock front.

Dessler [1961, 1962] discussed the stable features of the cavity surface and suggested that the geomagnetic K_p index may represent a measure of the time rate of change of the combination of plasma and magnetic pressure on the magnetosphere rather than the solar wind strength itself.

The shape of the magnetospheric boundary has been the subject of intensive theoretical study by a number of workers including Beard [1960, 1962a, 1962b], Hurley [1961], Midgley and Davis [1962], Slutz [1962], Spreiter and Briggs [1962a, 1962b], and Mead [1962]. But the idealizations and approximations made in these studies probably limit their application to the equatorial to moderate latitudes. The shape of the magnetospheric boundary in high latitudes and over the poles remains unexplored both experimentally and theoretically.

The sudden commencement of a magnetic storm can be interpreted as the effect of the impact of a solar plasma stream upon the magnetosphere. Dessler, Francis, and Parker [1960] interpreted the relatively slow buildup of sudden commencements, as indicated by their rise time of several minutes, to be caused by the cumulative effect of

hydromagnetic waves generated by the abrupt solar plasma impact arriving with varying transit times from different positions on the magnetospheric boundary to a point on the Earth's surface.

Wilson and Sugiura [1961] made an extensive morphological study of the sudden commencement of magnetic storms and presented a model for the sudden commencement. According to their model, the solar plasma impact on the magnetosphere generates compressional hydromagnetic waves which propagate to the Earth in low latitudes; while traveling through the outer regions of the magnetosphere the compressional hydromagnetic waves generate transverse hydromagnetic waves that propagate along the lines of magnetic force to the polar regions in the northern and southern hemispheres.

On the basis of the distribution of oppositely polarized sudden commencements Wilson [1962] determined the streaming direction of the storm producing solar plasma.

Akasofu and Chapman [1960] reviewed the studies made on the sudden commencement of magnetic storms and suggested that part of the sudden commencement field variations originates in the polar regions.

Matsushita [1960] analyzed the sudden commencement of magnetic storms observed during the IGY using data from the U.S. magnetic stations. Matsushita [1961] investigated sudden commencements, sudden impulses, and storm durations.

Vestine and Kern [1962] attributed the preliminary reverse impulse of the storm sudden commencement to the effect of electric currents driven in the ionosphere by a charge distribution in the polar regions; they discussed how such a charge distribution may be produced by distortions of the magnetosphere by a solar stream.

From investigations of a ring current that would produce the decrease of the magnetic field during the main phase of a magnetic storm, Akasofu [1960a] analyzed two large magnetic storms that occurred during the IGY.

Akasofu and Chapman [1961] proposed a model radiation belt responsible for the ring current; in their view protons of energies of the order of a few hundred kev are intermittently captured between 5 and 8 Earth radii during a magnetic storm, forming a transient belt. Akasofu, Cain, and Chapman [1961] computed the magnetic field decrease due to the proposed ring current.

Beard [1962b] and Akasofu [1962] discussed self-consistent calculations of the ring current.

Dessler, Hanson, and Parker [1961] suggested that hydromagnetic waves generated by the impact of solar plasma on the geomagnetic field may form shock waves in the magnetosphere which may heat the ambient protons to produce a diamagnetic ring current.

The observations by satellites and space probes have so far been limited to high-energy particles, and direct observational verification of the ring current has not yet been made.

Akasofu and Chapman [1963] examined the development of the main phase of magnetic storms by studying individual storms and showed that there are different types of magnetic storms. More studies of individual storms are needed to supplement studies of the average morphology of magnetic storms, as made by Sugiura and Chapman [1960].

Akasofu [1960b] studied large-scale auroral motions and polar magnetic disturbances, referring to the magnetic and auroral events of September 23, 1957.

Akasofu and Chapman [1962] discussed the ring current and the aurora on the basis of their neutral line discharge theory.

Kern [1961] discussed the distortion of the geomagnetic field by a solar stream and suggested that this distortion may produce auroral electrojets.

Investigation of geomagnetic micropulsations has greatly expanded during the period covered by this report. Workers at Lockheed Aircraft Corporation, National Bureau of Standards, University of Alaska, University of Texas, Air Force Cambridge Research Laboratories, U.S. Army Signal Research and Development Laboratory, California Institute of Technology, Lamont Geological Observatory, and others have contributed to progress in this field.

Campbell and his collaborators studied micropulsation activity in Alaska in relation to other disturbance phenomena [Campbell and Leinbach, 1961; Campbell and Rees, 1961; Campbell, 1962].

Tepley [1961] and Tepley and Wentworth [1962] observed hydromagnetic emissions, and Wentworth and Tepley [1962] discussed a possible mechanism for hydromagnetic emissions by electron bunches.

Sugiura [1961] presented examples of low-frequency transverse hydromagnetic waves generated in the magnetosphere and transmitted to the auroral zones along the lines of magnetic force.

Phenomena observed simultaneously at magnetically conjugate areas have attracted much attention in the past few years. Using the records from Macquarie Island and the Alaskan IGY stations, Wescott [1961] showed that in magnetically conjugate areas magnetic variations are remarkably similar. Mather and Wescott [1962] investigated Earth current records from a New Zealand-Alaska conjugate pair and found that major disturbances occur simultaneously.

Magnetic observations at a pair of conjugate areas have also been conducted by the National Bureau of Standards; their stations were in Eights, Antarctica, and in Quebec, Canada. Campbell and his collaborators made simultaneous observations of micropulsations at College and Macquarie Island. Other experiments are now being conducted.

In conclusion, the research activities in geomagnetic disturbances have considerably expanded in scope during the past three years, and many fruitful studies have been made with good results.

REFERENCES

Akasofu, S.-I., The ring current and the outer atmosphere, *J. Geophys. Res.*, **65,** 535–543, 1960*a*.

———, Large-scale auroral motions and polar magnetic disturbances, *J. Atmospheric Terrest. Phys.*, **19,** 10–25, 1960*b*.

———, On a self-consistent calculation of the ring current field, *J. Geophys. Res.*, **67,** 3617–3618, 1962.

———, Cain, J. C., and Chapman, S., The magnetic field of a model radiation belt, numerically computed, *J. Geophys. Res.*, **66,** 4013–4026, 1961.

———, and Chapman, S., The sudden commencement of geomagnetic storms, VRANIA, No. 250, 1960.

———, and ———, The ring current, geomagnetic disturbance, and the Van Allen belts, *J. Geophys. Res.*, **66,** 1321–1350, 1961.

———, and ———, The ring current and a neutral line discharge theory of the aurora polaris, *J. Phys. Soc. Japan*, **17,** Suppl. A–1, 169–173, 1962.

———, and ———, The development of the main phase of magnetic storms, *J. Geophys. Res.*, **68,** 125–129, 1963.

Beard, D. B., The interaction of the terrestrial magnetic field with the solar corpuscular radiation, *J. Geophys. Res.*, **65,** 3559–3568, 1960.

———, The interaction of the terrestrial magnetic field with the solar corpuscular radiation. 2. Second-order approximation, *J. Geophys. Res.*, **67,** 477–483, 1962*a*.

———, Self-consistent calculation of the ring current, *J. Geophys. Res.*, **67,** 3615–3616, 1962*b*.

Cahill, L. J., and Amazeen, P. G., The boundary of the geomagnetic field, *J. Geophys. Res.*, **68,** 1835–1843, 1963.

Campbell, W. H., Some auroral zone disturbances at times of magnetic micropulsation storms, *J. Phys. Soc. Japan*, **17,** Suppl. A–1, 112–116, 1962.

———, and Leinbach, H., Ionospheric absorption at times of auroral and magnetic pulsations, *J. Geophys. Res.*, **66,** 25–34, 1961.

———, and Rees, M. H., A study of auroral coruscations, *J. Geophys. Res.*, **66,** 41–55, 1961.

Coleman, P. J., Jr., Sonett, C. P., and Davis, L., Jr., On the interplanetary magnetic storm: Pioneer 5, *J. Geophys. Res.*, **66,** 2043–2046, 1961.

———, ———, Judge, D. L., and Smith, E. J., *J. Geophys. Res.*, **65,** 1856–1857, 1960.

Dessler, A. J., The stability of the interface between the solar wind and the geomagnetic field, *J. Geophys. Res.*, **66,** 3587–3590, 1961.

———, Further comments on stability of interface between solar wind and geomagnetic field, *J. Geophys. Res.*, **67,** 4892–4894, 1962.

———, Francis, W. E., and Parker, E. N., Geomagnetic storm sudden-commencement rise times, *J. Geophys. Res.*, **65,** 2715–2719, 1960.

———, Hanson, W. B., and Parker, E. N., Formation of the geomagnetic storm main-phase ring current, *J. Geophys. Res.*, **66,** 3631–3637, 1961.

Heppner, J. P., Ness, N. F., Scearce, C. S., and Skillman, T. L., Explorer 10 magnetic field measurements, *J. Geophys. Res.*, **68,** 1–46, 1963.

Hurley, J., Interaction of a streaming plasma with the magnetic field of a two-dimensional dipole, *Phys. Fluids*, **4,** 854–859, 1961.

Kellogg, P. J., Flow of plasma around the Earth, *J. Geophys. Res.*, **67,** 3805–3811, 1962.

Kern, J. W., Solar-stream distortion of the geomagnetic field and polar electrojets, *J. Geophys. Res.*, **66,** 1290–1292, 1961.

Mather, K. B., and Wescott, E. M., Telluric currents of geomagnetically conjugate stations in the Aleutian Islands and New Zealand, *J. Geophys. Res.*, **67,** 4825–4831, 1962.

Matsushita, S., Studies on sudden commencements of geomagnetic storms using IGY data from United States stations, *J. Geophys. Res.*, **65,** 1423–1435, 1960.

———, Sudden commencements, sudden impulses, and storm durations, *J. Geophys. Res.*, **67,** 3753–3777, 1962.

MEAD, G. D., Numerical solutions to the Chapman-Ferraro problem (abstract), *Trans. Am. Geophys. Union*, **43,** 459, 1962.

MIDGLEY, J. E., and DAVIS, L., Jr., *J. Geophys. Res.*, **67,** 499–504, 1962.

NEUGEBAUER, M., and SNYDER, C. W., Mariner measurements on the solar wind, Paper presented at Second Western National Meeting of AGU, Stanford, December 27–29, 1962.

———, and ———, The mission of Mariner 2: Preliminary observations, *Science*, **138,** 1095–1096, 1962.

SLUTZ, R. J., The shape of the geomagnetic field boundary under uniform external pressure, *J. Geophys. Res.*, **67,** 505–513, 1962.

SONETT, C. P., SMITH, E. J., and SIMS, A. R., *Space Research, I, Proc. Intern. Space Sci. Symp., 1st, Nice,* 1960, edited by H. Kallmann-Bijl, pp. 921–937, North-Holland Publishing Company, Amsterdam, 1960.

SPREITER, J. R., and BRIGGS, B. R., Theoretical determination of the form of the boundary of the solar corpuscular stream produced by interaction with the magnetic dipole field of the Earth, *J. Geophys. Res.*, **67,** 37–51, 1962*a*.

———, and ———, On the choice of condition to apply at the boundary of the geomagnetic field in the steady-state Chapman-Ferraro problem, *J. Geophys. Res.*, **67,** 2983–2985, 1962*b*.

SUGIURA, M., Evidence of low-frequency hydromagnetic waves in the exosphere, *J. Geophys. Res.*, **66,** 4087–4095, 1961.

———, and CHAPMAN, S., The average morphology of geomagnetic storms with sudden commencement, *Abhandl. Akad. Wiss. Gottingen, Math. Physik. Kl., Sonderheft,* **4,** 53, 1960.

TEPLEY, L. R., Observations of hydromagnetic emissions, *J. Geophys. Res.*, **66,** 1651–1658, 1961.

———, and WENTWORTH, R. C., Hydromagnetic emissions, X-ray bursts, and electron bunches, 1, Experimental results, *J. Geophys. Res.*, **67,** 3317–3333, 1962.

VESTINE, E. H., and KERN, J. W., Cause of the preliminary reverse impulse of storms, *J. Geophys. Res.*, **67,** 2181–2188, 1962.

WENTWORTH, R. C., and TEPLEY, L. R., Hydromagnetic emissions, X-ray bursts, and electron bunches, 2, Theoretical interpretation, *J. Geophys. Res.*, **67,** 3335–3343, 1962.

WESCOTT, E. M., Magnetic variations at conjugate points, *J. Geophys. Res.*, **66,** 1789–1792, 1961.

WILSON, C. R., Sudden commencement hydromagnetic waves and the enhanced solar wind direction, *J. Geophys. Res.*, **67,** 2054–2056, 1962.

———, and SUGUIRA, M., Hydromagnetic interpretation of sudden commencements of magnetic storms, *J. Geophys. Res.*, **66,** 4097–4111, 1961.

OSCILLATION OF THE GEOMAGNETIC FIELD LINES AND ASSOCIATED MAGNETIC PERTURBATIONS AT CONJUGATE POINTS

MASAHISA SUGIURA*

Goddard Space Flight Center

CHARLES R. WILSON

Geophysical Institute, University of Alaska

By means of an idealized elastic string model, the symmetry relations in the magnetic perturbation at a pair of conjugate points due to oscillation of the lines of force of the geomagnetic field are studied. For oscillations of odd modes the change in $H(\Delta H)$ is parallel, and that in the east component (ΔD) is antiparallel, at the conjugate points, whereas for oscillations of even modes ΔH is antiparallel and ΔD is parallel at the conjugate points. Hence in the former case the polarizations (as viewed downward) at the conjugate points are mirror images of each other with respect to the meridian plane, and in the latter case they are mirror images with respect to the equator. Several examples indicating these symmetry relations are presented.

1. INTRODUCTION

Resonant oscillations of the lines of force of the geomagnetic field have been discussed by a number of authors to explain geomagnetic pulsations [Dungey, 1954; Kato and Watanabe, 1956, 1957, 1958; Kato and Akasofu, 1956; Jacobs and Sinno, 1960]. If these pulsations are due to oscillations of the lines of force, they must be observed simultaneously at magnetically conjugate points. Sugiura [1961], Wilson and Sugiura [1961], Wilson [1963], Nagata et al. [1963], and Wescott and Mather [1963] have shown examples of simultaneous magnetic oscillations at conjugate points. However, this simultaneity of the oscillation alone does not prove the reality of the oscillation of the lines of force involving their entire length.

In this paper we present certain characteristics that must be demonstrated by pulsations at a pair of conjugate points if the pulsations are due to oscillation of the lines of force. An idealized model is used to derive the characteristics.

Observational data to which our discussions can be applied are limited, and we merely present here several examples to substantiate our idea.

2. ELASTIC STRING MODEL

Alfvén [1950] showed an analogy between the oscillation of a line of magnetic force and that of an elastic string. We will first examine the oscillation of an elastic string of uniform mass and uniform tension. Let the z axis be taken along the string, and let the string be fixed at $z=\pm a$. The string oscillates in the x-z plane, and the displacement of the string from its equilibrium position $(x=0)$ at time t is expressed by $x(z, t)$.

Then the equation for the oscillation is given by

$$\partial^2 x/\partial t^2 = V^2\, \partial^2 x/\partial z^2 \qquad (1)$$

where V^2 is the ratio of tension to the mass per unit length.

*NASA—National Academy of Sciences—National Research Council Senior Postdoctoral Resident Associate on leave of absence from Geophysical Institute, University of Alaska.

The solutions satisfying the boundary conditions for the fixed ends can be divided into two groups: in one group x is an even function of z: in the other, x is an odd function of z. Namely, the solution for the mth mode is of the form

$$\begin{aligned} &\cos\alpha_m z \text{ for } m=1, 3, 5, \cdots \\ &\quad x^{(m)}(z, t)=A_m \sin(\alpha_m V t+\phi_m) \\ &\sin\alpha_m z \text{ for } m=2, 4, 6, \cdots \end{aligned} \tag{2}$$

where A_m and ϕ_m are the amplitude and phase for the mth mode and where $\alpha_m=m\pi/(2a)$.

We now apply the above discussion to infinitesimal oscillation of the lines of force of a uniform magnetic field bounded by perfect conductors so that each line of force corresponds to the elastic string in the above illustration. The displacement $x(z, t)$ of a line of force from its equilibrium position can be described by (2), if V is taken to be the Alfvén velocity. The plasma between the conductors is assumed to be uniform, and hence V is a constant.

At any fixed t the equation of the line of force is $dx/b_x=dz/B_0$, where B_0 is the magnitude of the unperturbed field and b_x is the x component of the perturbation. Hence at time t

$$b_x=B_0\,\partial x/\partial z \tag{3}$$

It is noted that the magnetic perturbation is proportional to the derivative of the displacement and not to the displacement itself. Thus b_x varies with z as

$$b_x \propto \begin{cases} \sin\alpha_m z & \text{for } m=1, 3, 5, \cdots \\ \cos\alpha_m z & \text{for } m=2, 4, 6, \cdots \end{cases}$$

Now b_x is an odd function of z for odd modes, and an even function of z for even modes; the relation is opposite to the case of the displacement x.

For odd modes the midpoint ($z=0$) is a nodal point for b_x, whereas the magnitude of the displacement ($|x|$) is a maximum there. For even modes the situation is reversed. At the two ends, i.e., $z=\pm a$, the sign of b_x is opposite for odd modes modes and the same for even modes.

3. APPLICATION TO GEOMAGNETIC FIELD LINES

Although a straightforward application of the foregoing discussion to the oscillation of the geomagnetic field lines may appear to be an oversimplification, the above model may suffice to obtain qualitative geometrical characteristics of the oscillation of the lines of force.

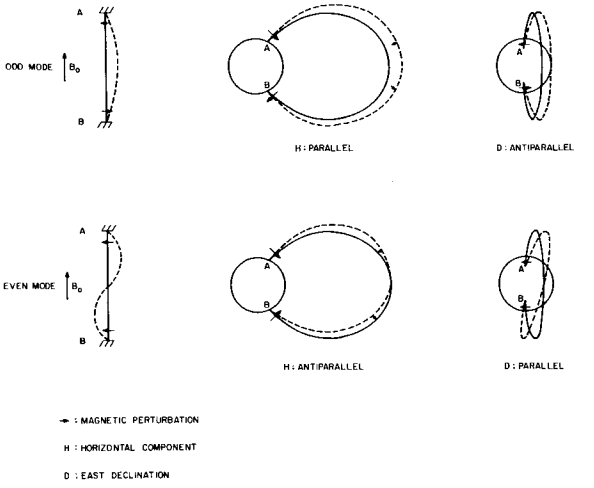


FIGURE 1.—The symmetry relations at magnetically conjugate points for oscillation of the lines of magnetic force: H, horizontal component; D, east declination; arrow, magnetic perturbation.

The lines of force of the geomagnetic field are curved, and the plasma in the magnetosphere is not uniform. But as long as the distribution of the electron density and the field configuration are symmetric with respect to the equator we can draw the following conclusions on the symmetry relations at conjugate points.

There is a neutral atmosphere between the earth and the ionosphere. However, since we deal with waves whose wavelengths are much greater than the thickness of neutral atmosphere, the existence of the insulator is ignored.

We assume that the decussions in section 2 are valid even when the string is curved so that its shape is that of a line of force of the geomagnetic field. The two fixed ends of the string correspond to a pair of conjugate points on the earth's surface.

We first take the x–z plane in the magnetic meridian containing the line of force, the z axis being along the line of force. Then the displacement of the line of force due to an oscillation is symmetric with respect to the equator for odd modes and antisymmetric for even modes. The magnitude of the magnetic perturbation in symmetric with respect to the equator for all modes.

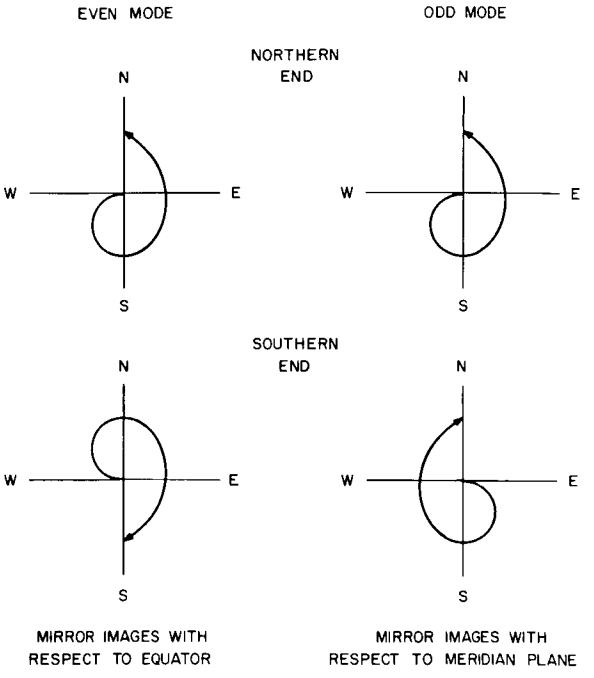


FIGURE 2.—Idealized symmetry relations in the polarization of magnetic perturbation at conjugate points due to oscillation of the lines of magnetic force.

But the direction of the magnetic perturbation (*inward or outward*) is opposite in the northern and southern hemispheres for odd modes and the same in both hemispheres for even modes. (See figure 1.)

Thus for oscillations of odd modes the perturbation in H is parallel at the conjugate points, whereas for oscillations of even modes it is antiparallel.

We now take the x–z plane in a curved surface containing the line of force and perpendicular to the meridian plane; then b_x may be considered the eastward component of the magnetic perturbation. The changes in this component are antiparallel at the conjugate points for odd modes and parallel for even modes.

If we combine the two components, we reach the following conclusion on the polarizations at conjugate points for elliptically polarized oscillation. We describe the polarization by the trace of the end point of the magnetic perturbation vector as viewed downward at ground level.

Then for oscillations of odd modes the polarizations at a pair of conjugate points are mirror images of each other with respect to the meridian plane, whereas for oscillations of even modes the polarizations at a pair of conjugate points are mirror images of each other with respect to the equatorial plane. Figure 2 illustrates these relations.

4. EXAMPLES OF TWO TYPES OF SYMMETRY

In our previous papers [Wilson and Sugiura, 1961, 1963: Wilson, 1963] we interpreted the sudden commencement of magnetic storms (ssc's) as a hydromagnetic phenomenon, and showed that in high latitudes the magnetic perturbation associated with an ssc can be considered as due to oscillation of the lines of force excited by the impact of solar plasma upon the magnetosphere. for the polarization analysis of ssc'c, rapid-run magnetograms are needed. However, the magnetic stations at College and Macquarie Island are the only conjugate pair operating rapid-run magnetometers in high latitudes. The Macquarie Island station began such a recording only recently, and at present rapid-run magnetograms from this station are available only for the three months from December 1962 to February 1963. During this period there were too few ssc's to test the conjugate relations proposed above. Hence we included three sudden impulses (si's) in addition to one ssc as examples frum the College–Macquarie Island pair. We treated si's and ssc's the same for our illustration, because the main

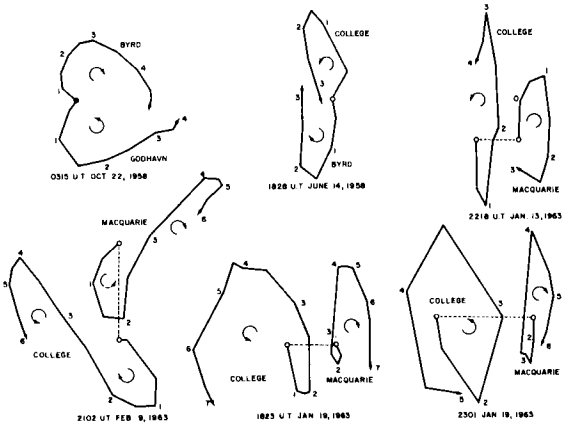


FIGURE 3.—Examples of two types of symmetry: the upper three correspond to oscillation of even mode; the lower three, to oscillation of odd mode.

morphological characteristics of si's are similar to those of ssc's [Matsushita, 1962; Nishida and Jacobs, 1962] (S. A. Bowling and C. R. Wilson, private communication, 1963).

We will now show examples of the two types of symmetry described in the preceding section by analyzing several ssc's and si's observed at a pair of conjugate stations, namely College and Macquarie Island, and at a few other stations in high latitudes.

The three vector diagrams in the upper half of figure 3 are examples of polarizations in the two hemispheres that are mirror images of each other with respect to the equator. According to the argument in section 3 they correspond to oscillation of an even mode.

The first two pairs, i.e., Byrd and Godhavn, and College and Byrd, are not conjugate, but the pairs lie in the same local-time zone in which the polarization pattern is the same [Wilson and Sugiura, 1961, 1963].

The three diagrams in the lower half of figure 3 are examples of polarizations that are mirror images at the conjugate points with respect to the meridian plane. They correspond to oscillation of an odd mode.

Of the six examples in figure 3 the two cases observed in 1958 and the one on February 9, 1963, are ssc's, and the remaining three are si's.

Because of the complexity of the phenomenon not all the ssc's and si's observed in high latitudes show clear symmetry relations at conjugate points. But the existence of such clear-cut cases as those shown above gives strong support to our view that ssc's and si's in high latitudes are mainly due to the oscillation of the lines of magnetic force.

Nagata et al. [1963] showed two vector diagrams for giant pulsations observed simultaneously at Syowa Base and Reykjavik, a pair of magnetically conjugate stations. It is interesting to note that in the example in their figure 2 the vector diagrams for the two stations in the horizontal plane are mirror images of each other with respect to the equator, and that the corresponding pair in their figure 3 gives vector diagrams that are mirror images with respect to the meridian plane. This suggests that the first example is an even harmonic and the second an odd harmonic. The period of the wave in the former is about 4.5 minutes and that of the latter about 8 minutes. (The latter period was determined from the Reykjavik record not reproduced in the paper by Nagata et al.) Thus these two cases may be example of oscillations of the first two modes, though the ration of the periods is not exactly 1 to 2.

Although much remains to be clarified as to the nature of the oscillation of the lines of magnetic force, the examples presented here seem to suggest that certain types of rapid geomagnetic variations are due to oscillations of the lines of force involving their entire length.

For any oscillation the displacement of a line of force can be expanded in a Fourier series. Then the magnetic perturbation at the end points of the line of force behaves differently for odd and even harmonics. The correlation coefficient taken for the perturbation at the conjugate points may, in idealized circumstances, be $+1$ or -1 according to the mode. Hence, in the study of magnetic perturbations at conjugate points, taking correlation coefficients in a straightforward manner may not be as meaningful as it might appear without due regard to the point expressed in this paper.

ACKNOWLEDGMENTS

We are grateful to the Bureau of Mineral Resources of Australia, the U. S. Coast and Geodetic Survey, and WDC-A for making the magnetic records available. We wish to thank Dr. J. P. Heppner for critically reading the manuscript.

This work was in part supported by the National Science Foundation under grant G-14734 and by the Air Force Cambridge Research Laboratories under contract AF 19(604)-7988.

REFERENCES

ALFVÉN, H., *Cosmical Electrodynamics*, Oxford University Press, New York, 1950.

DUNGEY, J. W., Electrodynamics of the outer atmosphere, *Penna. State Univ. Ionos. Res. Sci. Rept.*, 69, 1954.

JACOBS, J. A., and SINNO, K., World-wide characteristics of geomagnetic micropulsations, *Geophys. J.*, **3**, 333–353, 1960.

KATO, Y., and AKASOFU, S.-I., Outer atmospheric oscillation and geomagnetic micropulsations, *Sci. Rept. Tohoku Univ., Fifth Ser.*, **7**, 103–124, 1956.

———, and WATANABE, T., Further study on the cause of giant pulsations, *Sci. Rept. Tohoku Univ., Fifth Ser.*, **8,** 1–10, 1956.

———, and ———, Studies on geomagnetic pulsation *Pc, Sci. Rept. Tohoku Univ., Fifth Ser.*, **8,** 111–155, 1957.

———, and ———, Studies on geomagnetic storm in relation to geomagnetic pulsation, *J. Geophys. Res.*, **63,** 741–756, 1958.

MATSUSHITA, S., On geomagnetic sudden commencements, sudden impulses, and storm durations, *J. Geophys. Res.*, **67,** 3753–3777, 1962.

NAGATA, T., KOKUBUN, S., and IIJIMA, T., Geomagnetically conjugate relationships of giant pulsations at Syowa Base, Antarctica, and Reykjavik, Iceland, *J. Geophys. Res.*, **68,** 4621–4625, 1963.

NISHIDA, A., and JACOBS, J. A., World-wide changes in the geomagnetic field, *J. Geophys. Res.*, **67,** 525–540, 1962.

SUGIURA, M., Evidence of low-frequency hydromagnetic waves in the exosphere, *J. Geophys. Res.*, **66,** 4087–4095, 1961.

WESCOTT, E. M., and MATHER, K. B., Diurnal effect in magnetic conjugacy at very high latitude, *Nature*, **197,** 1259–1261, 1963.

WILSON, C. R., Ph.D. thesis, University of Alaska, 1963.

———, and SUGIURA, M., Hydromagnetic interpretation of sudden commencements of magnetic storms, *J. Geophys. Res.*, **66,** 4097–4111, 1961.

———, and ———, Discussion of our earlier paper, 'Hydromagnetic interpretation of sudden commencements of magnetic storms,' *J. Geophys. Res.*, 3314–3320, 1963.

TRAPPED ELECTRON TIME HISTORIES FOR L=1.18 TO L=1.30

J. A. WELCH, JR., AND R. L. KAUFMANN

Air Force Special Weapons Center

W. N. HESS

Goddard Space Flight Center

Numerical integration of a Fokker-Planck equation including atmospheric scattering and energy loss has been used to predict the time decay of electrons injected into the geomagnetic field by the July 9, 1962, nuclear explosion over Johnston Island. Calculations were performed for L= 1.18, 1.25, and 1.30 using an atmosphere characteristic of sunspot minimum and averaged over longitude, keeping L and B fixed. Deviations of the calculated and observed spatial distribution at one week after the injection are attributed to the overly simplified initial conditions. The early decay follows closely the form $\tau/(t+\tau)$ and is followed by an exponential decay at later times. For L=1.25, τ varies from 10^4 sec at B=0.21 to 10^3 sec at B=0.16 (the equator). The exponential decay for L=1.25 has a mean life of 1.3×10^3 seconds. Decay is accompanied by spectral hardening. After any substantial decay the spectrum is peaked at about 1.5 Mev.

INTRODUCTION

The loss of geomagnetically trapped electrons by atmospheric interactions has been studied for some time [Christofilos, 1959; Welch and Whitaker, 1959]. Wentworth et al. [1959] suggested that a Fokker–Planck formulation could be used for this study. A complete Fokker–Planck formulation including scattering and energy loss has been derived [MacDonald and Walt, 1961] and some results obtained [Walt and MacDonald, 1961]. These results hold for regions not too near the ends of the field lines and for steady state in the presence of a source.

We have used a formulation that is identical to that of Walt and MacDonald upon change in variables. The present work differs from theirs in that (1) it is valid for population changes deep in the atmosphere, i.e., near the ends of field lines, and (2) it follows the decay of an initial electron population with no additional source.

The atmospheric interaction is composed of many electron–atom or electron–ion collisions. Since, for the energy range of interest in artificial radiation belts from nuclear explosions (one-half Mev to several Mev), large-angle scattering leads to much longer lifetimes than multiple small-angle scattering, the large-angle scattering has been neglected. A small-angle scattering event can take place anywhere along the field line and will in general result in either a raising or a lowering of the electron's mirror point. At Mev energies a small energy loss accompanies scattering. The result of a large number of scattering events will be a diffusion of an electron along a field line and a gradual loss of energy from the electrons. The first step is to approximate this transport phenomenon with a Fokker–Planck expansion in powers of the mean scattering angle, which is small. Terms have been retained to the second power in this angle, and the first terms neglected contain the angle to the fourth power. To this approximation, terms involving scattering and terms involving energy loss appear separately. Details of

the Fokker–Planck expansion and of the solution to this equation are presented in a later section.

Results are presented for a simple initial distribution and are compared with recent observations [O'Brien et al., 1962; Hess, 1963] of trapped electrons injected by the Starfish nuclear explosion on July 9, 1962.

COORDINATE SYSTEM

In general we express the electron distribution function in terms of

γ, the electron's total energy in rest mass units.
B, the scalar magnetic field in gauss at the electron mirror point.
L, the McIlwain shell parameter [McIlwain, 1961] in earth radii.
ϕ, the east longitude.
t, time in seconds.

In this paper we will treat explicitly only conditions in which distributions are independent of ϕ and we can use atmospheric properties averaged over longitude.

Scattering sufficient to cause complete loss of an electron by changes in mirror-point field B will only cause a change in L on the order of a gyroradius ($\sim$1 kilometer). Since electron distributions of interest vary negligibly over such dimensions in L, we will neglect any diffusion in L. Populations in different L shells will of course have different time histories, but these can be calculated independently.

Diffusion in B will keep the mirror points within a shell of constant magnetic flux F, i.e., $B\times$(area perpendicular to B) = constant. Consequently, we have defined the distribution function as

$$U(L, B, \gamma, t)\, dF\, dB\, d\gamma = \text{the number of electrons at time } t \quad (1)$$

in $d\gamma$ at γ, in dB at B, and in the element of magnetic flux dF associated with the shell L. By defining the distribution function U in this way, we are dealing with the flux in a tube of force rather than along a line of force. For the record, $dF=(B_0/L^3)(d\phi\ L\ dL)(R_e)^2$ gauss cm^2, where $B_0=0.312$ gauss, the surface dipole equatorial field, and $R_e=6.371\times10^8$ cm is the radius of the earth. However, it is never necessary to relate dF to dL, as we shall see.

To relate U to m, the number of mirror points per cubic centimeter at any point, we set

$$U\, dB\, dF\, d\gamma = m\, dV\, d\gamma$$

where dV is an element of volume. Now we set $dF=B\ dA$, where dA is an element of area perpendicular to the field line, and further set $dV=dA\ ds$, where ds is an element of arc length along the field line. Thus we find

$$U=(m/B)(ds/dB)$$

The integral relations between U and the counting rate, and between an isotropic injection source and the resultant U, are developed in appendix 1.

THE FOKKER–PLANCK EQUATION

Under the assumption that changes in B and γ from individual collisions with atmospheric ions and atoms are small and uncorrelated, the general integral representation of the transport in B and γ can be expanded [MacDonald and Walt, 1961] to give

$$\partial U/\partial t = -(\partial/\partial B)(\langle\delta B\rangle U) - (\partial/\partial\gamma)(\langle\delta\gamma\rangle U) + \tfrac{1}{2}(\partial^2/\partial B^2)[\langle(\delta B)^2\rangle U] + \text{higher terms} \quad (2)$$

where the angle brackets denote the average change per unit time of the enclosed quantities. The evaluation of these angle brackets is outlined in appendix 2. Each term retained in (2) contains the square of the average scattering angle, and higher terms that have been neglected contain this angle to the fourth power. The major assumption is that changes in B and γ are small over a bounce period.

This equation is integrated numerically by an electronic computer. A line is divided into a number of cells of equal ΔB, the spectrum divided into a number of energy groups, and an initial mirror-point distribution is selected. Next, coefficients of the Fokker–Planck equation are calculated for each cell and for each energy group. This calculation involves an averaging of the effects of scattering events that take place at various points along a field line. Averaging is

done over a bounce period, so this calculation is not accurate when there is a large change during a single bounce. A model of the atmosphere must be introduced at this point in order to calculate coefficients of the Fokker–Planck equation. The model used is believed to be representative of average conditions near sunspot minimum, and is described in appendix 3. Air densities were averaged over longitude by following a particle as it drifts around the earth.

This should provide a reasonable average of the air density encountered by particles whose lifetimes are much longer than the drift period, but can lead to difficulties in the interpretation of low-altitude experimental measurements, as will be discussed later.

Once the coefficients of the Fokker–Planck equation have been calculated, a time step is selected so that the change in the mirror-point distribution will be small during this time interval. The mirror-point distribution is then calculated for each energy group at the end of the time interval using only scattering terms of the Fokker–Planck equation. Next the energy loss part of the equation is solved for this same time interval, and the fraction of particles in each cell and in each energy group that would have lost enough energy to be transferred to the next lower-energy group is determined. This process is continued, alternating between scattering and energy loss.

The boundary condition at the equator was $\partial U/\partial B=0$. The boundary condition at the bottom of the field line (that is, deep in the atmosphere) has a conceptual difficulty in that the assumption of small changes over a bounce period breaks down while U is still finite. Consequently we investigated the effects of several boundary conditions. For simplicity we first neglected energy loss and assumed the atmospheric density to be exponential in B,

$$\rho(B)=\rho_0 \exp(B/\Delta B) \tag{3}$$

Then we define a lifetime as

$$\tau(B)=\frac{\cos^2 \Delta a}{\frac{1}{2}\langle \epsilon^2 \rangle}=\frac{2\{1-[B/(B+\Delta B)]\}}{\langle \epsilon^2 \rangle} \tag{4}$$

where $\langle \epsilon^2 \rangle$ is the time averaged scattering angle factor from appendix 2, the factor $\frac{1}{2}$ corrects to projected angles, and Δa is the pitch angle at B of an electron whose mirror point is at $B+\Delta B$. From appendix 2 this lifetime is seen to be given approximately by

$$\tau(B)\approx \Delta B/|\langle \delta B \rangle| \tag{5}$$

which was used in a previous paper [Welch and Whitaker, 1959].

For a uniform initial population [$U(B, t=0)=$ constant], and a sink at the bottom [$U(B=B_c, t)=0$], we obtained

$$U(B, t)\rho(B)\simeq \text{constant} \tag{6}$$

for

$$t\gg\tau(B), \qquad B\lesssim B_c-\Delta B$$

and B not too near the equator. Replacing the sink at $B=B_c$ by the boundary condition,

$$(d/dB)(\ln U)=1/\Delta B \tag{7}$$

preserved the inverse density dependence right up to $B=B_c$. In another case we constructed an atmosphere as a sum of two exponentials with different scale heights,

$$\rho(B)=\rho_1 \exp(B/\Delta B_1)+\rho_2 \exp(B/\Delta B_2) \tag{8}$$

In this case we again obtained the inverse density dependence, even across the change in scale heights from ΔB_1 to ΔB_2. Further runs showed that these results are generally valid for any fairly uniform initial distribution. Since all these boundary conditions cause U to vary inversely with air density at low altitudes, the results of the calculation will not depend sensitively on which is actually used. In the present calculations a sink was placed at an altitude of 100 km over the South Atlantic in the region called the Brazilian anomaly.

It was mentioned previously that the longitude at which measurements are made becomes important at low altitudes. This effect has not been studied in detail yet, but a few test problems have been run to investigate the effect qualitatively. The results can be summarized by following a group of particles as they drift around the earth starting from a longitude over the Pacific Ocean. Initially there will be a fairly sharp cutoff at some value of B owing to the presence of the dense atmosphere. As the group drifts to the Atlantic, mirror points dip closer to the earth, and so the cutoff will move up

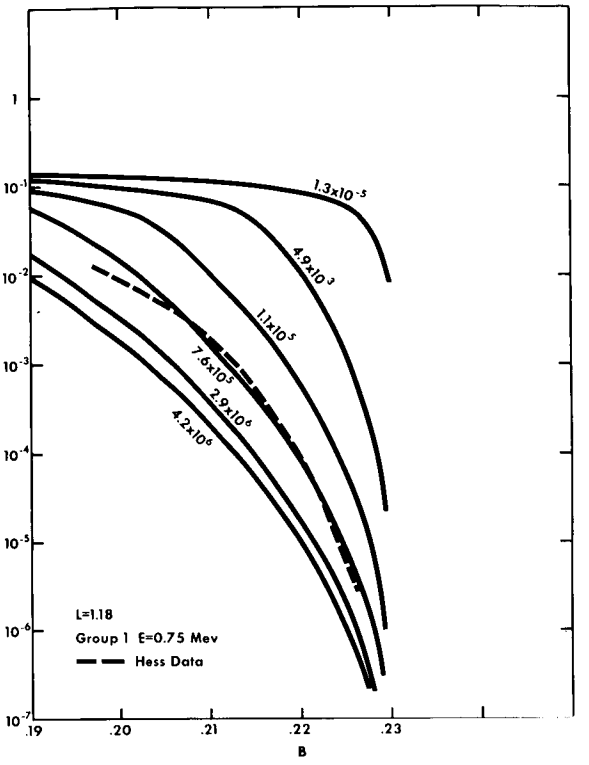


FIGURE 1.—Curves of the number density of 0.75-Mev electrons are plotted as a function of B on the $L=1.18$ field line. Each curve is labeled by the time after injection in seconds. Shown for comparison are data for one week after Starfish (Hess, 1963).

along the field line to a smaller value of B. As the group drifts back to the Pacific, particles will be scattered back down the field line so that the distribution will return approximately to its original shape. The effect of a longitude drift is, then, to cause the cutoff point to oscillate up and down along the field line. Above some altitude, only a small fraction of the particles will scatter appreciably during one drift period, so that the distribution will be nearly independent of longitude. Large changes will be observed at low altitudes, however, and it is important to specify longitude in low-altitude experimental measurements.

In comparing the results presented in the present work to experimental measurements, it must be recalled that averaged air densities are used. This means that the cutoff point and the scale height at low altitudes are determined primarily by the air density over the South Atlantic; experimental data taken near the anomaly should therefore agree fairly well with the results presented in this paper. Experimental measurements taken at other longitudes should show higher count rates at high field strengths.

The longitude effect is also very important in determining the actual place over the earth where most electrons are lost. This will be somewhere to the west of the Brazilian anomaly in the South Atlantic and will be a function of the energy of the trapped electron, but its exact position has not yet been determined.

RESULTS

The time history of electrons on the field lines characterized by $L=1.18$, 1.25, and 1.30 has been calculated starting with a fission β spectrum and a spatial distribution for U that is flat in B. This initial distribution was selected because it is the simplest distribution suggested by the satellite data, and a better estimate of the initial

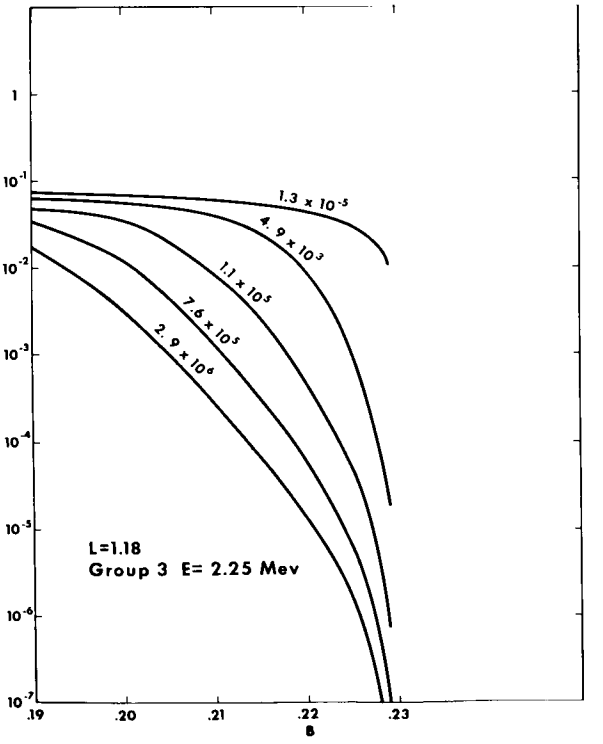


FIGURE 2.—Same as figure 1 for 2.25 Mev.

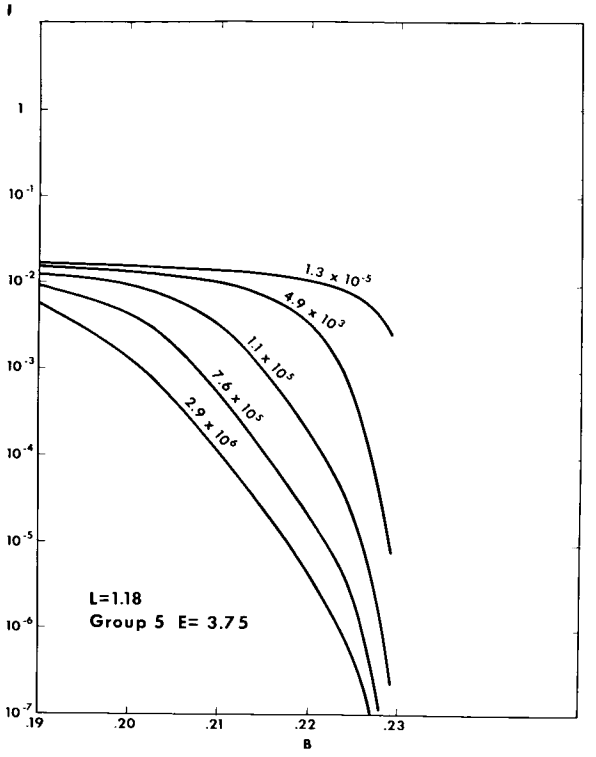


FIGURE 3.—Same as figure 1 for 3.75 Mev.

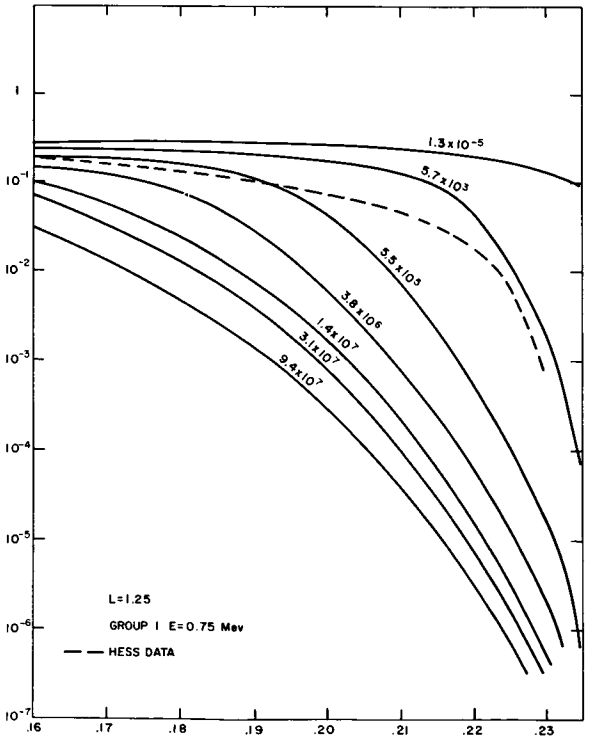


FIGURE 4.—Same as figure 1 for 0.75 Mev, $L=1.25$. Shown for comparison are data for 1 week after Starfish (Hess, 1963).

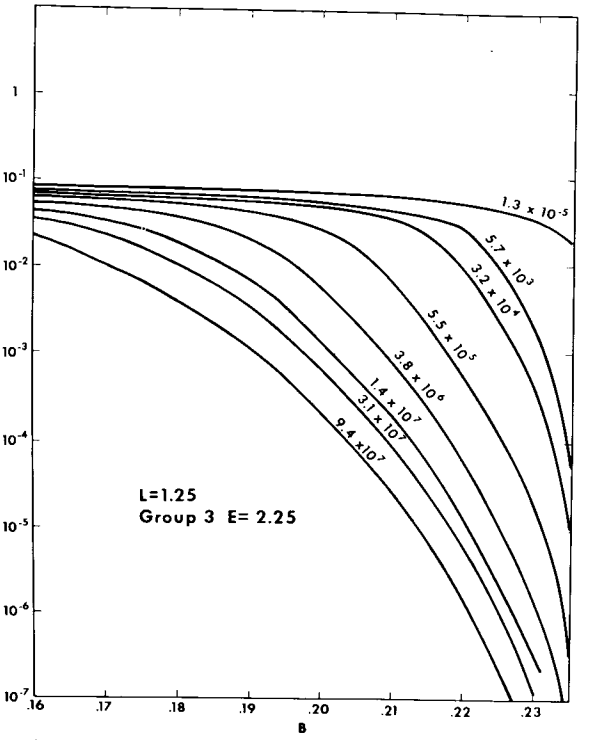


FIGURE 5.—Same as figure 1 for 2.25 Mev, $L=1.25$.

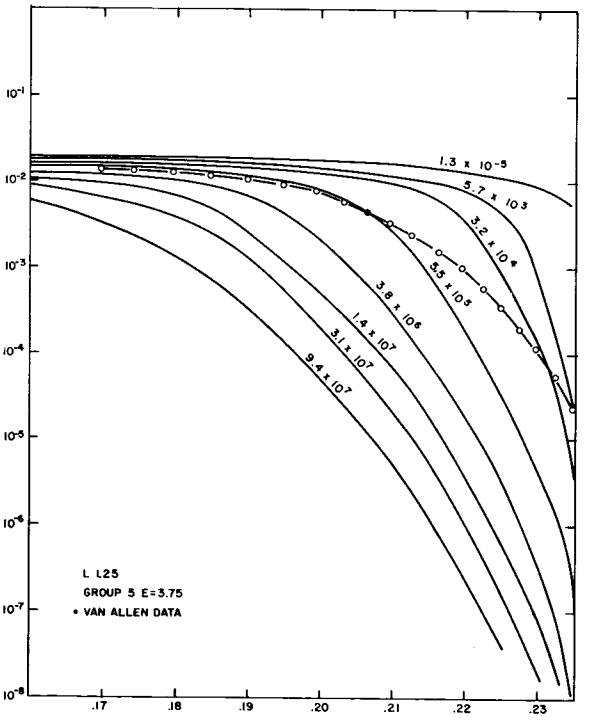


FIGURE 6.—Same as figure 1 for 3.75 Mev, $L=1.25$. Shown for comparison are data for 1 week after Starfish (O'Brien et al., 1962).

distribution produced by Starfish was not readily available.

The number density of electrons is plotted as a function of the field strength at various times for three energy groups and along three field lines in figures 1 to 9. Energy spectrums are plotted at various times at one particular point

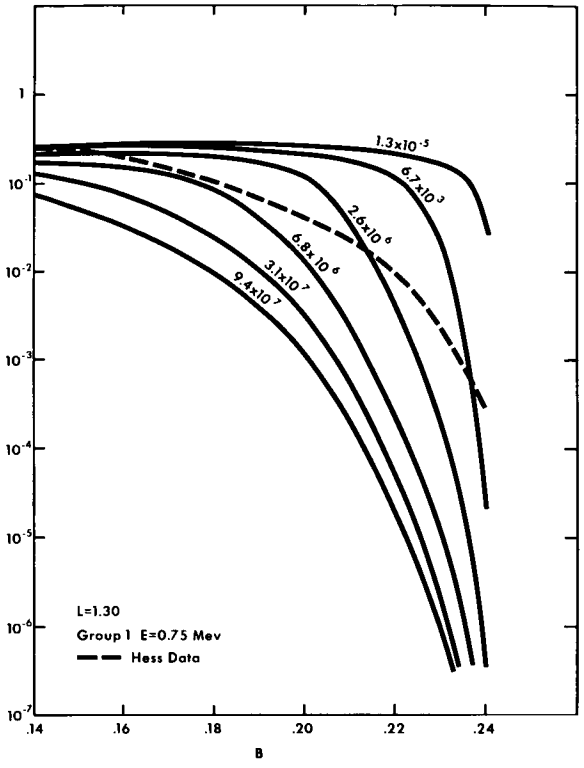


FIGURE 7.—Same as figure 1 for 0.75 Mev, $L=1.30$. Shown for comparison are data for 1 week after Starfish (Hess, 1963).

along each of the three field lines in figures 10 to 12. The particular point selected on each field line is somewhat off the equator, but plots made at a number of points along a field line all show a similar behavior. For comparison purposes, figure 13 is the same as figure 11 except that energy loss has been omitted. Figure 14 illustrates small changes in the spectrum along the $L=1.25$ field line better than figure 11 by showing the number density in several energy groups at a point somewhat off the equator as a function of time.

Several important features of the time decay

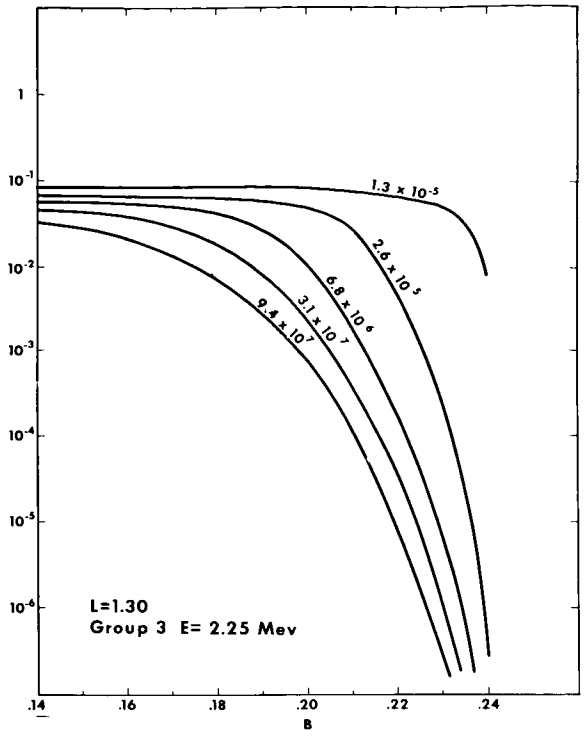


FIGURE 8.—Same as figure 1 for 2.25 Mev, $L=1.30$.

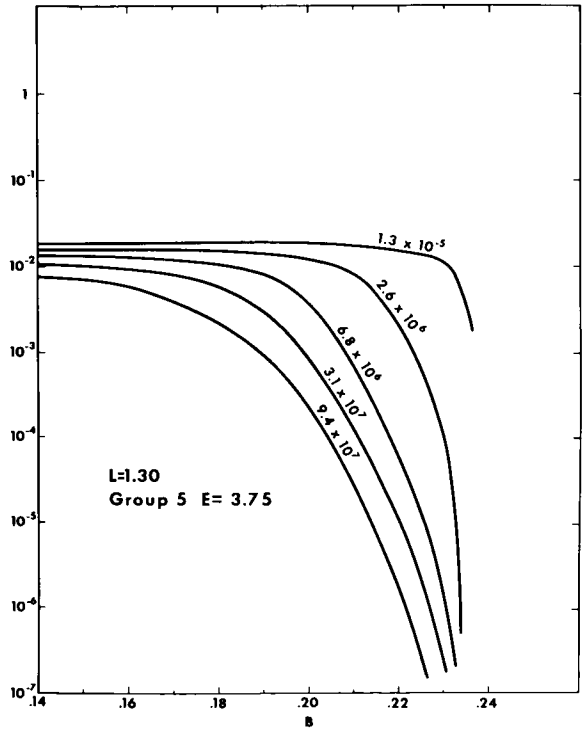


FIGURE 9.—Same as figure 1 for 3.75 Mev, $L=1.30$.

of electrons are evident from the figures. The first change that can be seen in figures 1 to 9 is a sharp drop in number densities at low altitudes (high B values). As time progresses, the point at which the number density drops sharply moves upward along the field line; the atmosphere therefore appears to cut off the initial distribution starting at low altitudes.

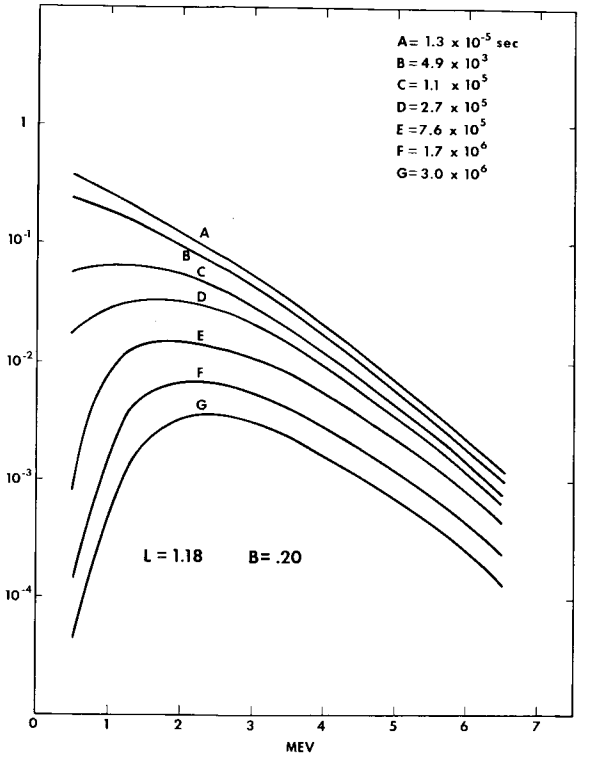


FIGURE 10.—Curves of the number density of electrons as a function of energy are plotted at several times for a point on the $L=1.18$ field line at which $B=0.20$.

Figures 10 to 12 show that the spectrum does change, particularly at early times in the low-energy region. It appears that the spectrum eventually reaches a stationary shape and then decays uniformly. Figure 14, however, illustrates small changes taking place in the energy spectrum more clearly, and it can be seen that some changes in the relative populations of the different energy groups are still taking place on the $L=1.25$ field line at three years. Eventually an equilibrium will be reached both in the energy spectrum and in the distribution along field lines, and

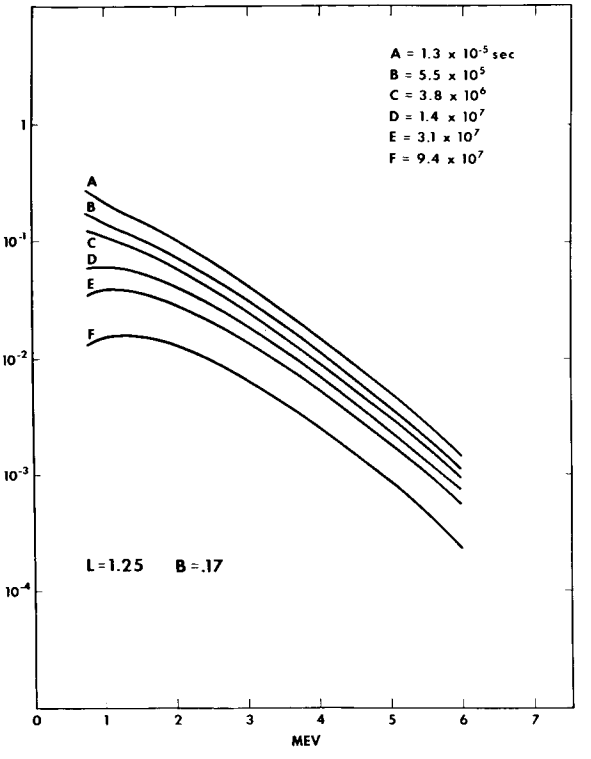


FIGURE 11.—Same as figure 10 for $L=1.25$, $B=0.17$.

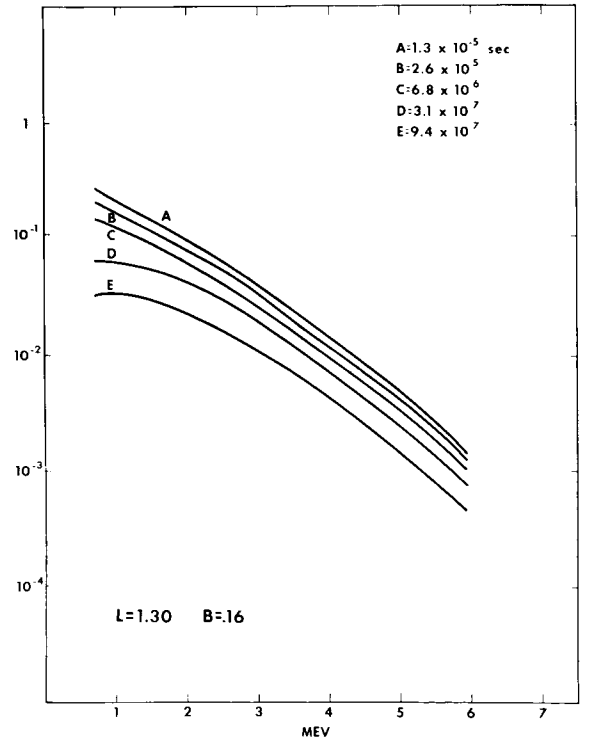


FIGURE 12.—Same as figure 10 for $L=1.30$, $B=0.16$.

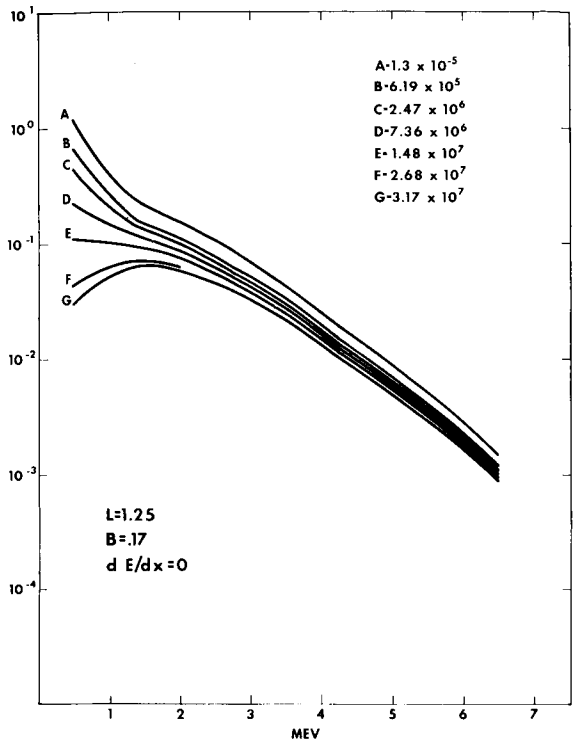


FIGURE 13.—Same as figure 11 except that energy loss has been neglected.

then the number densities will decay exponentially. This point will be discussed in more detail later.

One further general comment is important with regard to comparisons of problems run with and without energy loss. Without energy loss each energy group decays independently and identically except for a shift in the time scale. In this case we have a one-dimensional diffusion problem with mirror points moving only up and down along the field lines. By standard separation of variables we obtain the solution

$$U(B, \gamma, t) = F(B)\,T(t)\,G(\gamma) = \sum_i A_i(\gamma) f_i(B)\, \exp(-t/\tau_i)$$

where the A_i are determined by the initial distribution, f_i are the eigenfunctions, and τ_i are the eigenvalues ordered so that $\tau_{i+1} > \tau_i$ and are functions of γ.

For the assumed initial distribution having U constant in B, we find that A_1 is nonzero. Consequently at late times ($t \gg \tau_2$) we expect the first eigenfunction to dominate:

$$U \rightarrow A_1 f_1 \exp(-t/\tau_1)$$

When energy loss is included, we are dealing with a two-dimensional diffusion problem involving diffusion along the field line and also in energy. In this case, the part of the solution depending only on time can still be separated, but not the parts depending on B and γ. That is, we obtain

$$U(B, \gamma, t) = H(B, \gamma)\,T(t) = \sum_i C_i h_i(B, \gamma)\, \exp(-t/\tau_i)$$

where C_i are constants determined by the initial distribution, and τ_i are the eigenvalues ordered as before. They are not functions of γ but do depend on the range of γ, i.e., the boundary conditions in γ. This coupling of the energy loss and scattering implies that, even if only the first eigenfunction is present, the spectral shape will be a function of B. At late times ($t \gg \tau_2$) the decay will again be exponential with a fixed

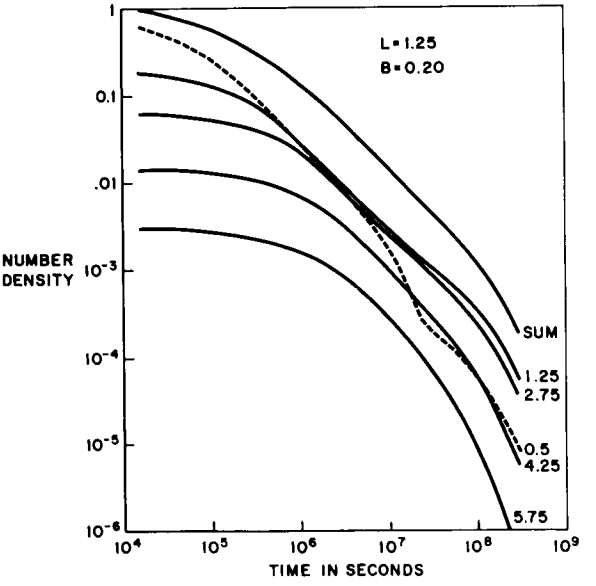


FIGURE 14.—Number density is plotted as a function of time for several energy groups at a point where $B = 0.20$ gauss on the $L = 1.25$ field line. The sum curve refers to the sum of all electrons with energies greater than 0.37 Mev.

spectral shape at any given B. For early times ($t<\tau_1$) the decay is governed by higher eigenfunctions and is consequently faster. If many eigenfunctions are needed to describe the distribution initially then that eigenfunction having $\tau_i \cong t$ will tend to govern the decay at time t. Higher eigenfunctions will have already decayed and lower ones are not yet decaying.

Details of the Spectral Changes

The effects of the coupling between the changes in energy spectrum and changes in the spatial distribution can be investigated by comparing figures 11 and 13. In figure 13 energy loss is neglected, and the decay is seen to be considerably different in character. The decay is slower, and an equilibrium energy spectrum is not ever produced. If this calculation had run long enough, the low-energy groups would be completely depopulated. It is clear that the lowest-energy group in figure 11 has an appreciable population only because it is being fed by energy loss from the next higher-energy group.

The extent of the spectral changes to be expected can be determined at any point along a field line for the 0.75-, 2.25-, and 3.75-Mev energy groups from figures 1 to 9. All points along the field line behave qualitatively the same, but the time scales will vary along the line. The most important general conclusion that can be drawn from the present calculations about spectral changes is that the spectrum will drop off rather sharply below 1.5 Mev, but the relative spectral change between higher energies is small.

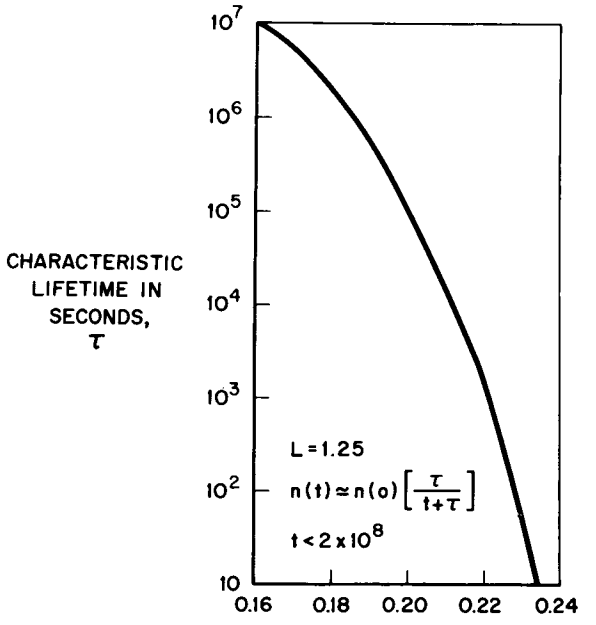


FIGURE 15.—Characteristic lifetimes are plotted as a function of B along the $L=1.25$ field line. This is good only for the total electron population, not for separate groups.

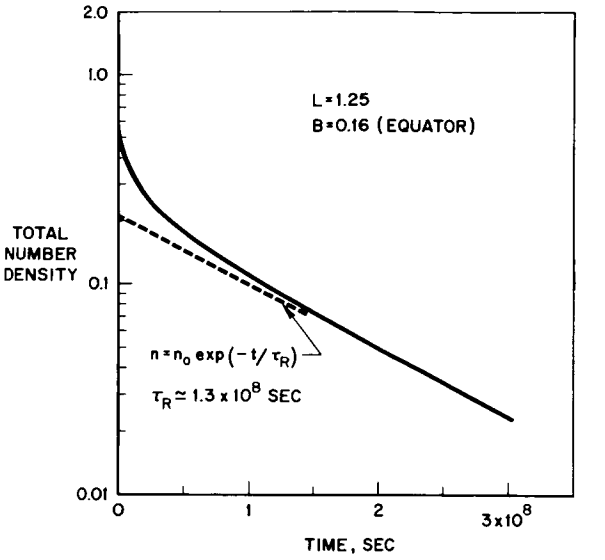


FIGURE 16.—Number density summed over all energies at the equator of the $L=1.25$ line vs. time. Note the exponential decay at late times.

Time Decay

A number of curves similar to figure 14 have been made at various points along the $L=1.25$ field line, and in all these the sum of electron number densities for all electrons of energies above 0.5 Mev can be approximated by the expression

$$n(B, t)=n(B, 0)[\tau/(t+\tau)] \qquad \tau=\tau(B)$$

where t is time, and τ is a characteristic lifetime. The parameter τ varies along the field line, and is plotted in figure 15. This expression fits the electron number density up to times of about 2×10^8 seconds (seven years) within 30 percent. As can be seen from figure 14, each individual energy group does not follow this decay law even though the sum over all energy groups does.

At later times the entire population decays exponentially with a characteristic time of about 1.3×10^8 seconds (see figure 16). All groups decay exponentially with the same characteristic time at late times, as was discussed above. In some cases a true exponential decay will never be observed experimentally because changes in atmospheric density will produce changes in the lifetime. During the period from 1962 to 1965 (near sunspot minimum) the atmospheric density should remain fairly near the distribution used in the present work, but decay at times greater than about three years after the Starfish explosion will be seriously influenced by changes in the composition of the atmosphere.

Comparison with Experiment

The results of these calculations are compared with what is known of the spacial distribution of flux resulting from the Starfish explosion. In figure 6 are shown the Injun results [O'Brien et al., 1962], and in figures 1, 4, and 7 are shown the composite results of Hess [1963]. These results have been normalized at the smallest value of B for each graph. The Injun results span some period of time, but are probably representative of $t\approx3\times10^5$ sec. Since the Injun detector counts electrons via the bremsstrahlung process, its energy is not well defined but is a few Mev average. The composite results in figures 1, 4, and 7 are for one week after Starfish, or about 6×10^5 sec. Since these results are for all fission electrons, they are most easily compared with the group 1 electrons centered at 0.75 Mev. Nearly half the fission electrons are in this energy interval.

The comparison of experiment with calculations for the $L=1.18$ line is quite good. For the $L=1.25$ and $L=1.30$ lines experimentally there are more electrons at large values of B than the calculations give. This could be due to a combination of three effects: (1) incorrect atmosphere, (2) longitude effect, and (3) the assumed initial mirror-point distribution being incorrect.

Although the atmosphere may be in error, we do not feel this is the source of the discrepancy. As was mentioned previously the calculations presented in this work are expected to underestimate count rates at low altitudes (high field strengths) at all longitudes except near the Brazilian anomaly. In particular, our calculations give zero count rate for B values beyond the cutoff, whereas there will be some electrons present everywhere except over the anomaly. Since many Injun 1 data are based on measurements taken over the United States (i.e., just before mirror points dip into the anomaly) the longitude effect would be maximized here. We have attempted to make numerical estimates of this effect and do not believe it can account for all the discrepancy noted between $B=0.20$ and $B=0.23$ on the $L=1.25$ field line. A firm conclusion will have to await a more detailed calculation of the effect of longitude.

In the event that the longitude effect is not sufficient to account for the experimental results, it must be concluded that the assumed initial mirror-point density was incorrect. It should be possible by trial and error to arrive at the initial mirror-point distribution required to fit experimental data.

SUMMARY

General features of the decay of an electron flux have been described in the present work. Both spectral and temporal changes have been discussed, and graphs were presented to illustrate this decay on field lines characterized by $L=1.18$, 1.25, and 1.30. The results of these calculations were then compared to experimental data taken one week after the Starfish explosion, and the observed discrepancies were discussed. To explain these discrepancies and to test the calculations in more detail, several further steps should be taken:

1. The longitude effect must be investigated in more detail before the results of the present calculations can be compared to experimental measurements made at low altitudes. This work is presently being carried out, and results should be available presently.

2. Other initial distributions should be investigated. It should be possible to find an initial distribution that will yield the experimentally measured distribution at one week after the explosion.

3. The calculated time decay should be compared to experiment. This will be possible as soon as data taken several months after the explosion have been collected and analyzed.

Appendix 1

RELATIONSHIP BETWEEN Q, U, AND n

Initial values of U can be obtained for distributed isotropic injection as follows. Let Q be the number of electrons per energy interval per cubic centimeter injected isotropically at the point defined by (L, B_p, ϕ). where L, the McIlwain [1961] parameter, describes the radius vector to the field line at the equator from the locally fitted dipole; B_p is the field strength, and ϕ is longitude. Let ds be an element of arc length along the local field line and dA be an element of area perpendicular to the line. Then we have

$$(Q\, d\gamma\, ds\, dA\, d\Omega/4\pi) \tag{1}$$

as the number of particles injected in ds, in $d\gamma$, in dA, and into the solid angle $d\Omega$. The element $d\Omega$ is related to an element of mirror-point field B. Remembering that the pitch angles α and $\pi-\alpha$ will result in the same mirror point, we obtain

$$d\Omega = (2\pi)[2d(\cos\alpha)] \qquad 0 \leq \alpha \leq \pi/2 \tag{2}$$

But since we have

$$\sin^2\alpha = B_p/B \tag{3}$$

we obtain

$$d\,(\cos\alpha) = d[1-B_p/B]^{1/2} = \frac{B_p\, dB}{2B^2[1-B_p/B]^{1/2}} \tag{4}$$

Noting that $B_p\, dA = dF$ defines the flux tube for injection points corresponding to the volume element $dB\, dF$ into which we are injecting mirror points, we obtain

$$U(\gamma, L, B, \phi) = \frac{1}{2B^2}\int \frac{Q\, ds}{[1-B_p/B]^{1/2}} \tag{5}$$

where the integral is over all s between the point (L, B, ϕ) and its conjugate, and Q and B_p are functions of s. To obtain U averaged in longitude we simply replace Q by $\bar{Q}$

$$\bar{Q}(\gamma, L, B_p) \equiv \frac{1}{2\pi}\int_0^{2\pi} Q(\gamma, L, B_p, \phi)\, d\phi \tag{6}$$

The number density can be obtained from the values of U as follows. Consider the volume $ds\, dA$ of the flux tube $dF = B_p\, dA$. The particles mirroring in $dB\, dF$ below this volume number

$$dN = U\, dB\, dF = U\, dB(B_p\, dA) \tag{7}$$

and spend a fraction of time g in the volume $(ds\, dA)$, where

$$g = \frac{ds}{v'' T} \cong \frac{ds}{2L[1-B_p/B]^{1/2}} \tag{8}$$

where T is the half-period of bounce and is given closely by $2L/v_{\text{total}}$ for all latitudes.

Thus the number density is given by

$$n(\gamma, L, B_p) = \int \frac{g\, dN}{dV} = \int \frac{ds\, U\, dB\, dA B_p}{2L[1-B_p/B]^{1/2}\, dA\, ds} \tag{9}$$

$$n(\gamma, L, B_p) = \frac{B_p}{2L}\int_{B_p}^{B^c} \frac{U\, dB}{[1-B_p/B]^{1/2}} \tag{10}$$

where $U = 0$ for $B > B_c$, the atmospheric cutoff.

Appendix 2

EVALUATION OF THE FOKKER–PLANCK COEFFICIENTS

The conservation of the electron magnetic moment leads to the mirror equation,

$$B = B_p/\sin^2\alpha \tag{1}$$

where B is the scalar field at the mirror point, and α is the pitch angle between the local velocity vector and the local magnetic field vector whose magnitude is B_p. Scattering of the electron by an atom will change α, and hence B, with B_p fixed. If ϵ is the angle between the old and new electron velocity vectors, and Ψ is the azimuthal angle of the scattering, we have for the new pitch angle α',

$$\cos\alpha' - \cos\alpha\cos\epsilon + \sin\alpha\sin\epsilon\sin\Psi \tag{2}$$

The corresponding change from B to B' is then obtained:

$$\begin{aligned}\delta B/B &= (B'-B)/B \\ &= (\sin^2\alpha/\sin^2\alpha') - 1 \\ &= (\cos^2\alpha' - \cos^2\alpha)/(1-\cos^2\alpha')\end{aligned} \tag{3}$$

Substituting (2) into (3), and collecting terms in powers of $\sin\epsilon$, we obtain

$$\frac{\delta B}{B} = \frac{2\sin\alpha\cos\alpha\sin\epsilon\cos\epsilon\sin\Psi + (\sin^2\alpha\sin^2\Psi - \cos^2\alpha)\sin^2\epsilon}{\sin^2\alpha - 2\sin\alpha\cos\alpha\sin\epsilon\cos\epsilon\sin\Psi + (\cos^2\alpha - \sin^2\alpha\sin^2\Psi)\sin^2\epsilon} \tag{4}$$

Performing long division and anticipating that ϵ is small we obtain

$$\begin{aligned}\delta B/B = {} & [2\cot\alpha\cos\epsilon\sin\Psi]\sin\epsilon \\ & + [\cot^2\alpha(4\sin^2\Psi - 1) + \sin^2\Psi]\sin^2\epsilon \\ & + 0(\sin\Psi\sin^3\epsilon) + 0(\sin^4\epsilon)\end{aligned} \tag{5}$$

Since physically all Ψ are equally probable and uncorrelated with ϵ, we obtain for an average over Ψ

$$\langle \delta B \rangle/B = (\cot^2\alpha + \tfrac{1}{2})\sin^2\epsilon + 0(\sin^4\epsilon) \tag{6}$$

Bu squaring (5) and again averaging over Ψ we obtain

$$\langle (\delta B)^2 \rangle/B^2 = (2\cot^2\alpha)\sin^2\epsilon + 0(\sin^4\epsilon) \tag{7}$$

Averages over ϵ are obtained by means of a screened Coulomb scattering cross section from Moliere [1948]. Remembering that we wish averages of the time rate of change of δB and $(\delta B)^2$, we construct:

$$\langle \sin^2\epsilon \rangle \equiv \sum_i N_i(B) v \int_0^\pi \sin^2\epsilon (d\sigma_i/d\epsilon)\, d\epsilon \tag{8}$$

where $N_i(B)$ is the longitude-average (see appendix 3) number density of the ith atmospheric component (e.g., oxygen, nitrogen, helium), v is the total velocity, and $(d\sigma/d\epsilon)$ is the differential cross section [Moliere, 1948]

$$d\sigma_i/d\epsilon = \frac{8\pi\sin\epsilon Z^2 e^4}{m_0 v^4 \gamma^2}\left[\frac{0.10}{\epsilon^2 + 9(\eta^2/\gamma^2)} + \frac{0.55}{\epsilon^2 + 0.36(\eta^2/\gamma^2)} + \frac{0.35}{\epsilon^2 + 0.022(\eta^2/\gamma^2)}\right]^2$$

where e is the electronic charge, m_0 is the electron rest mass, and $\eta = 0.56 Z^{1/3}(2\pi e^2/hv)$. In this equation Z, the Hartree–Fock value of the atomic number, equals 5.8 for oxygen and 1.4 for helium.

For energies above 500 kev we obtain the approximate numerical result,

$$\langle \sin^2\epsilon \rangle = [(1.2+0.2\ \ln\gamma)/\gamma^2]\times 10^{-11} \times [N_O+0.08N_{He}]\ \text{sec}^{-1} \quad (9)$$

for an atmosphere consisting of N_{He} (helium atoms)/cc and N_O (oxygen and nitrogen atoms)/cc.

After substituting (9) into (6) and (7), averages were taken over a bounce period. These were performed numerically using the atmospheric model of appendix 3 after weighing the residence time at each value of B by $(\cos\alpha)^{-1}$.

The energy loss coefficient $\langle \delta\gamma \rangle$ was evaluated for the region $\gamma > 2$ by assuming minimum ionization

$$\rho^{-1}\,(d\gamma/dx) = (\rho v)^{-1}\,(d\gamma/dt) = 4 \text{ per } (\text{g/cm}^2) \quad (10)$$

This results in

$$\langle \delta\gamma \rangle = 3.2\times 10^{-12}(\bar{N}_O+0.25\bar{N}_{He})\ \text{sec}^{-1} \quad (11)$$

where $\bar{N}_O$ and $\bar{N}_{He}$ denote weighted averages over the bounce period.

Appendix 3

THE ATMOSPHERIC MODEL

One of the largest uncertainties in this analysis involves the values of the atmospheric densities. The proper value to be used in this analysis is a diurnal average density for the present part of the solar cycle. There are few good experimental data on high-altitude densities.

The only complete theoretical study of the solar cycle and diurnal atmospheric density and composition is that of Harris and Priester [1962]. This study assumes that the intensity of the ultraviolet upper atmosphere heat source is proportional to the 10-cm solar radiation flux (denoted by S and given in units of 10^{-22} watt/m^2 cps). This has some theoretical basis [Waldmeier and Müller, 1950] and it seems to be true empirically, at least over a certain part of the solar cycle, judging from a study of the 27-day variations [Priester, 1959; Jacchia, 1959]. Harris and Priester also assumed a second heat source with a forenoon maximum in order to avoid a diurnal maximum temperature at 5 p.m. local time, which would otherwise occur. The observed diurnal maximum is at about 2 p.m. This second heat source is assumed to vary in the same way with the monthly averaged 10-cm solar flux as the ultraviolet source. Since the position of the diurnal maximum is not of importance to us here (we are interested only in diurnal averaged densities), the physical realness of this second heat source is unimportant.

The Harris and Priester atmosphere also includes helium, which is known now to be a major component at high altitudes. We will not consider hydrogen here because it is a principle atmospheric constituent only at very high altitudes, and also it is a very inefficient coulomb scatterer because of its low Z.

According to current information on the solar flux ($S \cong 100$), the Harris and Priester diurnal averaged atmosphere is as shown in table 1.

We are neglecting ions in this atmosphere. Although one ion is about twice as efficient as one atom in scattering, because of the increased range of its force, the percentage ionization is quite small. Table 1 was used as the atmospheric model.

Let us see what checks there are on the accuracy of this theoretical model atmosphere.

First, the presence of helium at high altitude, which was predicted some time ago [Nicolet, 1961], has been demonstrated experimentally

TABLE 1

Altitude, km	n(N)	n(O)	n(He)
120	1.2×10^{12}	3.2×10^{11}	2.5×10^{7}
200	7.2×10^{10}	3.7×10^{9}	5.9×10^{6}
300	2.0×10^{8}	3.7×10^{8}	3.2×10^{6}
400	7.5×10^{6}	5.5×10^{7}	2.0×10^{6}
500	3.2×10^{5}	9.0×10^{6}	1.3×10^{6}
600	1.5×10^{4}	1.6×10^{6}	8.2×10^{5}
800	40	5.6×10^{4}	3.5×10^{5}
1000	—	2.4×10^{3}	1.6×10^{5}
1200	—	120	7.5×10^{4}
1500	—	2	2.6×10^{4}
2000	—	—	5.3×10^{3}

[Hansen, 1962]. The experiment by Hale [1961] in October 1960 that gave this information was carried out at about 1500 local time. This gave, at 2000 km, a temperature of 1600°K and $n(\mathrm{O}) = 2\times10^{3}$ cm^{-3} and $n(\mathrm{He}) = 1.5\times10^{5}$ cm^{-3}. The Harris and Priester atmosphere for $S=200$ would give $n(\mathrm{O}) = 1.2\times10^{3}$ and $n(\mathrm{He}) = 8\times10^{4}$, in good agreement. Their atmosphere here gives $T=1750$°K, somewhat higher than 1600°K but probably inside the limits of experimental error.

Bourdeau et al. [1962] also found helium on the Explorer 8 satellite. A Langmuir probe gave an

electron temperature of $1750 \pm 200°$ at 0615 local time at 1630 km on November 24, 1960. This was a disturbed period. A retarding potential ion trap gave a similar ion temperature and a ratio of $He^+/O^+ = 1.3$ at 1630 km. An analysis of this experiment showed rather good agreement with the Hansen atmosphere.

Another experiment of Sharp et al. [1962] involving a microphone gage gave a density of $2.1 \pm 0.4 \times 10^{-15}$ g/cm^2 at 500 km during the day and $2.6 \pm 1.3 \times 10^{-16}$ g/cm^2 at 550 km at night on January 31, 1961. This agrees well with Harris and Priester's data for $S = 150$.

Serbu et al. [1961] measured electron temperatures on the Explorer 8 satellite in November 1960 and got a diurnal variation of from about 1100°K to about 1800°K, in quite good agreement with Harris and Priester for $S = 200$. More recent measurements by L. H. Brace (private communication, 1962) gave temperatures of 1700°K at noon in March 1961 and 1200°K at midnight in December 1961. These fit the Harris and Priester atmosphere for $S = 150$ moderately well.

In June 1962 an electron temperature measurement on a Scout Rocket (J. L. Donley, private communication, 1962) gave a temperature of $T = 800$°K at 0100 local time. Harris and Priester gave $T = 720$°K for $S = 100$ and $T = 940$°K for $S = 150$.

According to recent measurements of the 10.7-cm solar flux (W. Covington, private communication, 1962), we actually are very nearly at $S = 100$ now.

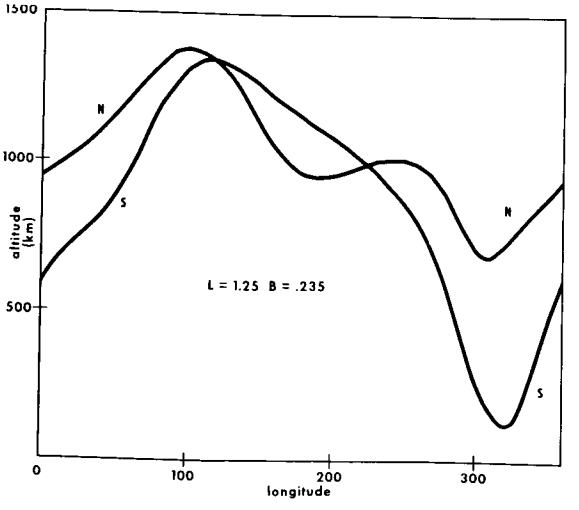


FIGURE 17.—The variation of altitude with longitude of a ring of constant B and L in both the northern and southern hemispheres.

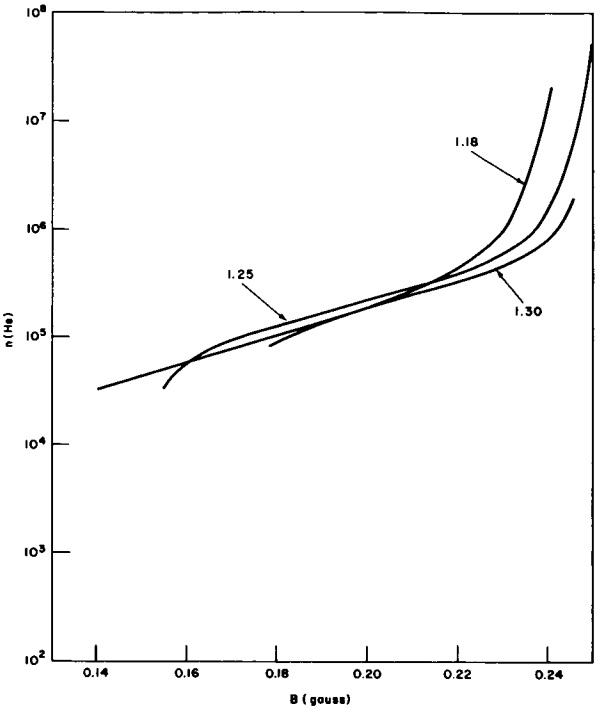


FIGURE 18.—The atmospheric helium density averaged in longitude and local time near sunspot minimum.

Another source of information about upper atmosphere densities comes from studying satellite drag. From the observations on the change of period of a satellite, we can calculate the air density ρ.

Jacchia and Slowey [see Sharp et al., 1962], studying satellite $1958\beta_2$, found that in February 1961 the atmospheric density at 2300 local time is $\rho = 1.5 \times 10^{-17}$ g/cm^2 at 710 km and at 1400 LT is $\rho = 8 \times 10^{-17}$ at the same altitude. These agree quite well with Harris and Priester for $S = 150$. From studying the drag on the Echo satellite, R. Bryant (private communication, 1962) has found that for July 1961 $\bar{\rho} = 1 \times 10^{-18}$ g/cm^3 and for July 1962 $\bar{\rho} = 0.5 \times 10^{-18}$ at 1400 km. This corresponds for July 1961, from Harris and Priester, to $S = 150$ and for July 1962 to $S \sim 110$.

The Harris and Priester model atmosphere is only good for low latitudes. Recent data from the Ariel Satellite [Bourdeau, 1962] show that above 20° latitude the exospheric temperatures

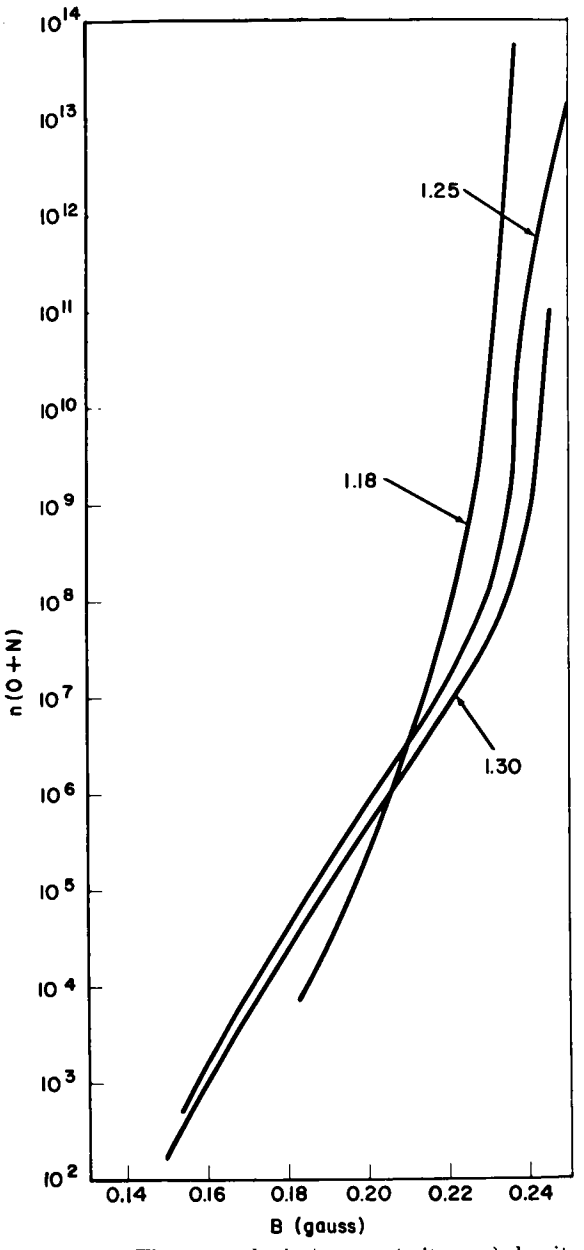


FIGURE 19.—The atmospheric (oxygen+nitrogen) density averaged in longitude and local time near sunspot minimum.

increase a good deal. The agreement of Ariel data with Harris and Priester's data at low altitude is quite good. Since we are only interested here in $\lambda < 20°$, the Harris and Priester atmosphere will be adequate.

On the basis of reasonably good agreement between experiments and the Harris and Priester model atmosphere, we will use the $S=100$ diurnal average atmosphere given in table 1. The value of $S=100$ seems quite good now. It should decrease slowly for the next two years, but not change very much in total during that time.

The Harris and Priester diurnal average atmosphere given in table 1 must be averaged in longitude to be used in the calculations. This is done by selecting a value for B and L and calculating the corresponding altitude. This is repeated at 5° intervals in longitude. One such longitude-altitude graph is shown in figure 17. (The Jensen and Cain [1960] coefficients of the magnetic field were used in this calculation.) By interpolating in table 1, the values of the atmospheric densities n(He) and n(O+N) are obtained at all points in longitude and are averaged. The process is then repeated for a different B–L set. The result of this is the longitudinally averaged densities of n(He) and n(O+N) shown in figure 18 and figure 19. These are used in the calculations.

ACKNOWLEDGMENTS

Lt. J. W. H. Taylor and R. W. Murray of Air Force Special Weapons Center constructed the program for the problem and R. F. Baxter, and W. F. Cahill of Goddard Space Flight Center performed the calculations.

REFERENCES

BOURDEAU, R., Paper presented at the 9th Annual Meeting of the IRE–PGNS, the International Symposium on Space Phenomena and Measurement, Detroit, October 16, 1962.

BOURDEAU, R. E., WHIPPLE, E. C., Jr., DONLEY, J. L., and BAUER, S. J., Experimental evidence for the presence of helium ions based on Explorer 8 satellite data, *J. Geophys. Res.*, **67,** 467–477, 1962.

CHRISTOFILOS, N. C., The Argus experiment, *J. Geophys. Res.*, **64,** 869–876, 1959.

HALE, L. C., Ionospheric measurements with a multigrid potential analyzer (abstract, *J. Geophys. Res.*, **66,** 1554, 1961.

HANSEN, W. P., Upper atmospheric helium ions, *J. Geophys. Res.*, **67,** 183–189, 1962.

HARRIS, I., and PRIESTER, W., Theoretical models for the solar-cycle variation of the upper atmosphere, *NASA Tech. Note D*–1444, 1962.

HESS, W. N., The artificial radiation belt, *J. Geophys. Res.*, **68**(3), 667–683, 1963.

JACCHIA, L. G., *Harvard College Observ. Announcement Card* 1423, January 16, 1959.

———, Solar effects on the acceleration of artificial satellites, *Smithsonian Inst. Astrophys. Observ., Res. Space Sci. Spec. Rept.* 29, September 1959.

JENSEN, D. C., and CAIN, J. C., An interim geomagnetic field (abstract), *J. Geophys. Res.*, **67,** 3568, 1962.

———, MURRAY, R. W., and WELCH, J. A., Jr., *Tables of Adiabatic Invariants for the Geomagnetic Field* 1955.0, Air Force Special Weapons Center, Albuquerque, New Mexico, 1960.

MACDONALD, W. M., and WALT, M., Distribution function of magnetically confined electrons in a scattering atmosphere, *Ann. Phys.*, **15,** 44–62, 1961.

MCILWAIN, C. E., Coordinates for mapping the distribution of magnetically trapped particles, *J. Geophys. Res.*, **66,** 3681–3692, 1961.

MOLIERE, G., Electron scattering in solids, translated by H. S. W. Massey, *Advan. Electron.*, **4,** 1–68, 1952.

NICOLET, M., Helium, an important constituent in the lower exosphere, *J. Geophys. Res.*, **66,** 2263–2265, 1961.

O'BRIEN, B. J., LAUGHLIN, C. D., and VAN ALLEN, J. A,. Geomagnetically trapped radiation produced by a high-altitude nuclear explosion on July 9, 1962, *Nature*, **195,** 939, 1962.

PRIESTER, W., *Mitt. Univ. Sternwarte Bonn no.* 24, 1959; also *Naturwissenschaften*, **46**(6), 197–198, March 1959.

SERBU, G. P., BOURDEAU, R. E., and DONLEY, J. L., Electron temperature measurements on Explorer 8 satellite, *J. Geophys. Res.*, **66,** 4313–4316, 1961.

SHARP, G. W., HANSON, W. B., and MCKIBBEN, D. D., Atmospheric density measurements with a satellite-borne microphone gage, *J. Geophys. Res.*, **67,** 1375–1382, 1962.

WALDMEIER, M., and MÜLLER, H., Die sonnenstrahlung im gebiet von $\lambda = 10$ cm, *Z. Astrophys.*, **27**(1), 58–72, April 5, 1950.

WALT, M., and MACDONALD, W. M., Energy spectrum of electrons trapped in the geomagnetic field, *J. Geophys. Res.*, **66,** 2047–2052, 1961.

WELCH, J. A., Jr., and WHITAKER, W. A., Theory of geomagnetically trapped electrons from an artificial source, *J. Geophys. Res.*, **64,** 909–922, 1959.

WENTWORTH, R. C., MACDONALD, W. M., and SINGER, S. F., Lifetimes of trapped radiation belt particles determined by Coulomb scattering, *Phys. Fluids*, **2,** 499, 1959.

DISCUSSION OF OUR EARLIER PAPER 'HYDROMAGNETIC INTERPRETATION OF SUDDEN COMMENCEMENTS OF MAGNETIC STORMS'

CHARLES R. WILSON

Geophysical Institute, College, Alaska

MASAHISA SUGIURA*

Goddard Space Flight Center

On page 3771 of his recent paper in this Journal Matsushita [1962] discussed the disagreement of his analysis with ours [Wilson and Sugiura, 1961] regarding the polarization rules of the elliptically polarized hydromagnetic waves in the magnetic storm sudden commencements. The purpose of this communication is to correct the misunderstanding that may have resulted from Matsushita's criticism, which appears to be based on an analysis much less extensive and complete than ours. The discussions given below are based on the results of analysis of about six hundred and fifty vector diagrams of the SC horizontal perturbation ΔH.

The material for this analysis was drawn from virtually all the rapid-run magnetograms available at the IGY World Data Center A; the analysis concerns 93 SC's that occurred during the period beginning with the IGY and ending September 1961. Though for a few of these SC's data from only a few magnetic observatories were available, records from seven stations, on the average, were scaled for each SC. In addition, data were collected from 18 to 30 magnetic observatories (out of the 38 observatories listed in table 2) for ten selected SC's that are evenly distributed over the Greenwich day.

The polarization rules were derived and tested in the following way. First, polarization characteristics were determined for 9 to 67 SC's for 14 observatories (listed in table 1) and the results were all combined to derive rules by which the polarizations are governed. Secondly, polarization characteristics were examined for 18 to 30 observatories (table 2) for each of 10 SC's to test whether or not the rules obtained by the first method are obeyed by individual SC's.

These analyses confirmed our previous findings and gave further support to our interpretation of SC's. The general picture of our interpretation of SC's can be summarized as follows.

The primary effect of the impact of a solar gas cloud on the surface of the magnetosphere is to create a longitudinal hydromagnetic shock wave. Since near the equatorial plane the magnetic field is nearly perpendicular to the equatorial plane, the longitudinal hydromagnetic wave propagates toward the earth essentially in this mode. Because of the curvature of the lines of magnetic force, the longitudinal hydromagnetic wave generates transverse hydromagnetic waves by coupling. These transverse hydromagnetic waves are transmitted to the earth essentially along the lines of magnetic force. The coupling is most effective in the outer regions of the magnetosphere where the magnetic field is weak; from these regions the transverse hydromagnetic waves propagate to the earth in high latitudes.

The above picture is greatly idealized; in reality the magnetic perturbation created by the impact of the solar plasma will propagate in a more

*National Academy of Sciences—National Research Council fellow on leave of absence from the Geophysical Institute, College, Alaska.

TABLE 1.—*Percentage of SC's That are Elliptically Polarized, and Percentage of SC's (Among these Elliptically Polarized SC's) Obeying the Polarization Rules*

1	2	3	4	5		6	
Magnetic observatory	Geomagnetic latitude	R_e, earth radii†	Total number of SC's, N	Elliptically polarized SC's		SC's obeying polarization rules	
				Number, N_e	$(N_e/N)\times 100\%$	Number, N_e^*	$(N_e^*/N_e)\times 100\%$
	(°)				%		%
Murchison Bay	75.25N	15.5	20	15	75	7	47
College	64.7N	5.4	55	32	58	26	81
Healy	63.6N	5.0	31	22	71	21	96
Sitka	60.0N	4.0	67	52	78	50	96
Lovö	58.2N	3.6	30	14	47	12	86
Fredericksburg	49.6N	2.4	62	23	37	21	91
Tucson	40.4N	1.7	36	11	30	10	91
Honolulu	21.0N	1.14	37	8	21	6	75
Guam	3.9N	1.005	9	0	0	—	—
Apia	16.0S	1.08	9	2	22	1	50
Hermanus	33.3S	1.45	33	16	48	13	81
Watheroo	41.7S	1.78	34	16	47	11	69
Byrd Station	70.6S	9.0	31	20	65	18	90
Wilkes Station	77.8S	22.5	47	36	77	24	66
Total			501	267		220	

†R_e is the geocentric distance of the point at which the magnetic field line (of the unperturbed dipole) through the station crosses the equatorial plane.

complex manner. However, we believe that the above mechanism gives the essential structure of the sudden commencement of a magnetic storm. The basis of this belief is the fact that this mechanism explains some of the major features of SC's observed on the earth's surface.

If an SC magnetic perturbation observed at an observatory were a pure transverse hydromagnetic wave, its polarization would be circular, whereas if it were a pure longitudinal hydromagnetic wave, its polarization would be linear. The polarization of a magnetic perturbation can be studied by tracing the end point of the magnetic perturbation vector. Only the projection of the magnetic vector onto the horizontal plane is investigated here.

The analysis of the characteristics of 650 such vector diagrams for the first few minutes of SC's has shown that the earth can be divided roughly into three zones according to the polarization of SC's. The polarization characteristics, the ranges in geomagnetic latitude Φ of these zones, and the ranges in geocentric distance R_e (in units of earth radius) of the regions in the magnetosphere which are the projections of the three zones onto the equatorial plane along the lines of magnetic force are:

(a) linear polarization for $0°\leq\Phi\leq 40°$ and $1.0\leq R_e\leq 1.7$.

(b) mixture of elliptical and linear polarizations for $40°\leq\Phi\leq 55°$ and $1.7<R_e\leq 3.0$.

(c) elliptical polarization for $55°<\Phi\leq 90°$ and $R_e>3.0$.

We interpret this result as follows: The SC magnetic perturbation primarily consists of longitudinal hydromagnetic waves in zone (a), and of transverse hydromagnetic waves in zone (c). Between these two zones there is a transitional zone where either one of these two modes or a mixed mode may be observed.

Table 1 summarizes the results of analysis

regarding (1) percentage of SC's that are elliptically polarized, and (2) among these elliptically polarized SC's, percentage of SC's that obey the polarization rules. It is clearly seen that the percentage of elliptically polarized SC's increases with increasing latitude. For the southern hemisphere rapid-run magnetic data are scarce; however, the results for the five observatories used do indicate the same trend.

The sense of rotation, either clockwise of counterclockwise *when viewed downward* on the earth's surface, of the circularly polarized SC's obeys the following two rules:

1. In each of the northern and southern hemispheres the sense of rotation is opposite in two quadrants separated by the meridian plane through 10 hours and 22 hours.
2. In each meridian plane the sense of rotation of magnetic vector is opposite in the northern and southern hemispheres.

In the northern hemisphere the rotation is

TABLE 2.—*Results of Analysis of Ten SC's*

1 Time of SC		2 Stations for which data were available	3 Stations at which SC was elliptically polarized	
UT	Date	Number, S	Number, S_e	$(S_e/S)\times 100\%$
0042	July 5,1957	23	17	74%
0315	Oct. 22,1958	26	23	88
0843	Sept. 3,1958	26	18	69
0930	Sept. 16,1958	29	24	83
1050	Jan. 25,1958	27	22	81
1300	Sept. 4,1957	18	16	89
1529	July 31,1958	30	26	87
1652	May 31,1958	22	17	77
1821	Nov. 6,1957	22	21	95
1920	Aug. 29,1957	24	18	75
		Total 247	Total 202	Mean 82

Stations (Geomagnetic Latitude)

1. Thule (N 89°0)
2. Godhavn (N 79°8)
3. Murchison Bay (N 75°3)
4. Reykjavik (N 70°2)
5. Point Barrow (N 68°6)
6. Tromsö (N 67°2)
7. Kiruna (N 65°3)
8. College (N 64°7)
9. Big Delta (N 64°4)
10. Healy (N 63°6)
11. Lerwick (N 62°5)
12. Sitka (N 60°0)
13. Eskdalemuir (N 58°4)
14. Lovö (N 58°2)
15. Valentia (N 56°7)
16. Rude Skov (N 55°9)
17. Hartland (N 54°6)
18. Manhay (N 52°0)
19. Dourbes (N 52°0)
20. Fredericksburg (N 49°6)
21. Ponta Delgada (N 45°6)
22. Toledo (N 43°9)
23. Tucson (N 40°4)
24. Memambetsu (N 34°1)
25. Kakioka (N 26°0)
26. Honolulu (N 21°0)
27. Paramaribo (N 17°0)
28. Guam (N 3°9)
29. Koror (S 3°3)
30. Hollandia (S 12°5)
31. Apia (S 16°0)
32. Hermanus (S 33°3)
33. Watheroo (S 41°7)
34. Byrd Station (S 70°6)
35. Little America (S 74°0)
36. Dumont d'Urville (S 75°5)
37. Wilkes (S 77°8)
38. Scott Base (S 79°0)

TABLE 2.—Continued

1		4		5		
Time of SC		Stations at which elliptically polarized SC obeyed the rules		Same as columns 3 and 4, excluding stations with $\Phi>72°(R_c>10.5)$		
UT	Date	Number, S_e*	$(S_e*/S_e)\times100\%$	$S_e'(S_e$ with $\Phi>72°)$	S_e'*	$(S_e'*/S_e')\times100\%$
0042	July 5,1957	12	70%	12	10	83%
0315	Oct. 22,1958	19	83	18	17	95
0843	Sept. 3,1958	13	72	14	11	99
0930	Sept. 16,1958	23	96	18	17	95
1050	Jan. 25,1958	16	73	16	11	69
1300	Sept. 4,1957	13	81	11	10	91
1529	July 31,1958	17	56	20	15	75
1652	May 31,1958	14	82	13	13	100
1821	Nov. 6,1957	13	62	17	11	65
1920	Aug. 29,1957	17	95	13	12	92
		Total 157	Mean 77	Total 152	Total 127	Mean 84

clockwise from 10 hours to 22 hours; this specifies the sense of rotation in the remaining three quadrants according to the two rules given above.

The two rules can be combined into one: the SC hydromagnetic perturbation is transmitted to the earth in the ordinary mode in the morning hemisphere (22 hours to 10 hours) and in the extraordinary mode in the afternoon hemisphere (10 hours to 22 hours). However, in our earlier paper we represented the mode by the sense of rotation of the magnetic vector; hence, for the sake of uniformity of representation this same way of expressing the mode of hydromagnetic waves is used here.

In the group of elliptically polarized SC's the percentage of cases in which the two rules are obeyed ranges from 91% to 69%, with a mean of 86% (column 6, table 1), excluding the very high latitude stations (Murchison Bay and Wilkes) and Apia (for which there were too few cases); the percentage of agreement with the polarization rules is independent of latitude. It is noteworthy that even for stations at which circularly polarized SC's are not very frequent, e.g., Fredericksburg, Tucson, and Honolulu, the majority of circularly polarized SC's follow the pattern set by the rules.

In the foregoing discussions the rules were derived statistically treating the time variable, 'local time,' at each station as being equivalent to the spatial variable, 'longitude with respect to the sun,' at a fixed moment of time. It is desirable, therefore, to test to what extent individual SC's follow the rules. For this purpose analysis was made on ten SC's (listed in table 2). The stations used are listed in table 2; the number of stations for which it was possible to analyze the data with an acceptable accuracy is shown in column 2.

For the ten SC's, the percentage of stations at which SC's were elliptically polarized ranges from 69% to 95% (column 3), indicating a high rate of occurrence of elliptically polarized SC's. Taking the cases in which SC is observed to be elliptically polarized, the number and the percentage of stations at which the polarization of SC was in agreement with that expected from the polarization rules are given in column 4. The agreement between the observed and the predicted pattern of SC polarization can be considered good. The average of the above percentage $(S_e*/S_e)\times100\%$ (column 4, table 2) is 77%; the corresponding percentage in the statistical analysis discussed earlier is 84% (column 6, table 1).

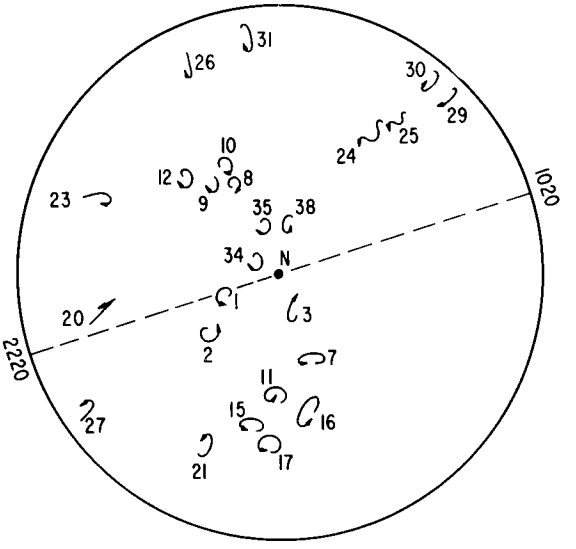
Recent plasma measurements by Mariner 2 indicate that *steady* solar wind exists in inter-

planetary space [Neugebauer and Snyder, 1962], and because of this continuous flow of solar plasma the magnetosphere is confined in a 'cavity.' Observational evidence for the 'cavity surface' has been provided by magnetic measurements made with instruments aboard Pioneer 1 [Sonnett et al., 1960], Pioneer 5 [Coleman et al., 1960], Explorer 10 [Heppner et al., 1962], and Explorer 12 [Cahill and Amazeen, 1963]. The measurement by Explorer 10 indicated that the cavity surface extends to distances greater than 20 earth radii on the dark side of the earth. On the day side of the earth the cavity surface is near 10 earth radii.

For an unperturbed centered dipole the magnetic field lines crossing the equator at a geocentric distance of 10 earth radii intersect the earth's surface at 72° latitude. Because of the magnetic field produced by the electric current on the boundary surface, the field lines will be distorted. Theoretical studies of the shape of the cavity surface have been made by a number of workers [Dungey, 1961, 1962; Hurley, 1961; Midgley and Davis, 1962; Slutz, 1962; Spreiter and Briggs, 1962; Mead, 1962]. However, the idealizations and approximations used in these studies make it difficult to determine the distortions of the magnetic field lines that are anchored in the earth near the magnetic poles. Magnetic measurement in the magnetosphere over the polar regions has not been made.

Thus it is not certain whether the magnetic field lines intersecting the earth's surface near the magnetic poles are all crossing the equator on the dark side of the earth, as has been suggested by Johnson [1960], or some of these lines of force are connected to those of interplanetary magnetic fields or to the disordered lines of force of the irregular magnetic field that has been observed by Pioneer 1, Pioneer 5, Explorer 10, Explorer 12, and Explorer 14 and has been discussed by Dessler [1962]. If the magnetic field lines originating near the magnetic poles are not interhemispherically connected, the modes, i.e., ordinary or extraordinary, of the transverse SC waves may not agree between high-latitude stations in the adjacent northern and southern quadrants. Even if these lines of force are connected on the dark side of the earth at distances greater than 10 earth radii, the magnetic field may be too weak for an efficient coupling of longitudinal and transverse hydromagnetic waves. The dipole field at 10 earth radii is about 30 γ, and the SC amplitudes are about 100 γ or even greater in high latitudes. Even if the focusing of the wave energy in the propagation toward the earth is taken into account, the amplitude of the SC perturbation is of the same order of magnitude as, or greater than, the unperturbed field. Then the coupling and transmission of hydromagnetic waves may be much more complicated than those expected from the linearized theory.

In any event there is reason to suppose that the SC polarization rules may not be satisfied in the regions near the magnetic poles. How large is the extent of these regions is as yet a matter of speculation. However, it is of interest to set some upper boundaries in latitude, one in each hemi-



0315 U.T. OCTOBER 22, 1958

FIGURE 1.—Map of the horizontal perturbation vector (ΔH) diagrams for the sudden commencement of 0315 U.T. October 22, 1958, for 20 magnetic observatories in the northern hemisphere. The vector diagrams are drawn at the projected positions of the stations on the equatorial plane. The numbers by each vector diagram refer to the station identification list in table 2. The meridian plane separating the two zones of opposite polarization (extraordinary mode 1020 to 2220 hours and ordinary mode from 2220 to 1020 hours) is shown.

sphere, above which the mode of transverse SC waves is at random.

If we set these boundaries at geomagnetic latitudes ±72° (as was indicated earlier, the lines of force crossing the earth's surface at 72° latitude intersect the equatorial plane at a geocentric distance of 10 earth radii) then, for the stations with latitude greater than 72°, on the average only 53% of the SC's obey the polarization rules, indicating that such boundaries as defined above appear to exist near 72°. In column 5 of table 2 the percentage of stations with latitude less than 72° for which elliptically polarized SC's obey the polarization rules is given; the average percentage is now increased to 84% from 77%, which is the corresponding percentage when all the stations are included (column 4). Though the difference is very slight, the percentage does increase by excluding the stations in the polar cap.

The vector diagrams for the first few minutes of the SC of November 22, 1958, at 0315 UT, for 20 stations in the northern hemisphere are shown in figure 1. The position of each station is projected onto the equatorial plane along the line perpendicular to this plane, and the vector diagram for the station is placed at the projected position of the station. The vector diagram is drawn as if the plane of the projection were the horizontal plane at the station, and the vector diagram is so oriented that the direction of the magnetic north for the vector diagram is pointed toward the center. The scales are not uniform for all the stations. The number by each vector diagram identifies the station according to the list of stations in table 2.

The SC is elliptically polarized and satisfies the polarization rules at 14 of the 20 stations. The meridian by which two modes are separated is approximately through 10 h 20 m and 22 h 20 m [Wilson, 1962].

At station 3, Murchison Bay, the sense of rotation is opposite to that expected from the

TABLE 3.—*Analysis of SC Vector Diagrams of Looped or Hooked Form*

1	2	3			4	5	6
Station	Geomagnetic latitude (deg)	Time of SC			Rotation expected from rules	Rotation in loop or hook, observed	Form of vector diagram
		Date	UT	LT			
Honolulu	21.0	Feb. 11,1958	0125	1425	C	CC	l
		Aug. 6,1957	0508	1808	C	C	h
		Oct. 28,1958	0650	1950	C	C	h
		July 8,1958	0748	2048	C	C	h
		July 2,1957	0857	2157	C	C	h
		Sept. 22,1957	1345	0245	CC	CC	h
Tucson	40.4	July 5,1957	0042	1742	C	C	h
		June 7,1958	0046	1746	C	C	h
		Sept. 16,1958	0930	0230	CC	CC	h
		Dec. 19,1957	0937	0237	CC	CC	h
		Aug. 29,1957	1920	1220	C	C	h
Fredericksburg	49.6	Sept. 29,1957	0016	1916	C	C	h
		July 7,1957	0042	1942	C	C	h
		Jan. 5,1959	0137	2037	C	C	h
		May 24,1959	0540	0040	CC	C	h
		Dec. 19,1957	0937	0437	CC	CC	h
		Jan. 29,1960	1937	1437	C	C	l
Lerwick	62.5	July 31,1958	1529	1529	C	C	l
Lovö	58.2	July 31,1958	1529	1629	C	C	l
Rude Skov	55.9	July 31,1958	1529	1629	C	C	l
Hartland	54.6	July 31,1958	1529	1529	C	C	l

rules; however, the line of magnetic force (of the unperturbed dipole) through this station crosses the equator at 15.5 earth radii.

At station 26, Honolulu, and at station 27, Paramaribo, the vector diagram exhibits a hooked form that can be interpreted as an indication of the initial arrival of a predominantly longitudinal wave followed by the arrival of a transverse wave. The sense of rotation of the magnetic vector in the hooks for stations 26 and 27 agrees with that expected from the polarization rules for higher latitudes.

These hooked forms frequently occur at middle to low latitudes on the side of the earth away from the sun. There is often a counterpart on the day side of the earth where the vector diagram has an initial loop followed by a linear increase in H. However, such a feature is not apparent in the example shown in figure 1.

To demonstrate that the polarization rules are obeyed by those SC's that exhibit a hook or an inverted hook in their vector diagrams, the sense of rotation of the magnetic vector for the hooked part is shown in table 3. Column 4 of table 3 indicates for each observatory SC the sense of rotation, clockwise (C) or counterclockwise (CC), expected from the quadrant polarization rules. The sense of rotation actually observed in the hooked part of each SC is shown in column 5. Agreement between columns 4 and 5 is good; 80% of the 17 observatory SC's listed in table 3 for Honolulu, Tucson, and Fredericksburg show agreement.

For the SC of July 21, 1958, at 1529 UT the four stations, Lerwick, Lovö, Rude Skov, and Hartland, were all on the day side of the earth. The vector diagrams for these stations showed a loop or inverted hook form, and the polarization characteristics were in agreement with those expected from the rules.

In column 6 of table 3 the form of the vector diagram in each SC is indicated; here l and h signify loop and hook forms, respectively.

With the data presented in this communication we believe that the polarization rules for the magnetic storm commencements are well founded.

More detailed discussions of our analysis will be presented later.

ACKNOWLEDGMENT

This project was supported by a grant from the National Science Foundation and by the Air Force Cambridge Research Laboratories through contract AF 19(604)-7988.

REFERENCES

CAHILL, L. J., and AMAZEEN, P. G., The boundary of the geomagnetic field, *J. Geophys. Res.*, **68**(7): April 1, 1963.

COLEMAN, P. J., Jr., SONETT, C. P., JUDGE, D. L., and SMITH, E. J., *J. Geophys. Res.*, **65,** 1856–1857, 1960.

DESSLER, A. J., *J. Geophys. Res.*, **67,** 4892–4894, 1962.

DUNGEY, J. W., *J. Geophys. Res.*, **66,** 1043–1047, 1961.

———, *J. Phys. Soc. Japan*, **17,** Suppl. A–2, 15–19, 1962.

HEPPNER, J. P., NESS, N. F., SCEARCE, C. S., and SKILLMAN, T. L., *J. Geophys. Res.*, **68,** 1–46, 1963.

HURLEY, J., *Phys. Fluids*, **4,** 854–859, 1961.

JOHNSON, F. S., *J. Geophys. Res.*, **65,** 3049–3051, 1960.

MATSUSHITA, S., *J. Geophys. Res.*, **67,** 3753–3777, 1962.

MEAD, G. D., Paper presented at the 2nd Western Natl. Meeting Am. Geophys. Union, Stanford, December 27–29, 1962.

MIDGELEY, J. E., and DAVIS, L., Jr., *J. Geophys. Res.*, **67,** 499–504, 1962.

NEUGEBAUER, M., and SNYDER, C. W., *Science*, **138,** 1095–1096, 1962; paper presented at the symposium on Results of Venus Probe, Mariner 2, 2nd Western Natl. Meeting Am. Geophys. Union, Stanford, December 27–29, 1962.

SLUTZ, R. J., *J. Geophys. Res.*, **67,** 505–513, 1962.

SONETT, C. P., SMITH, E. J., and SIMS, A. R., *Proc. Intern. Space Sci. Symp., 1st, Nice*, pp. 921–937, 1960.

SPREITER, J. R., and BRIGGS, B. R., *J. Geophys. Res.*, **67,** 37–51 and 2983–2985, 1962.

WILSON, C. R., and SUGIURA, M., *J. Geophys. Res.*, **66,** 4097–4111, 1961.

———, *J. Geophys. Res.*, **67,** 2054–2056, 1962.

HYDROMAGNETIC WAVES GENERATED BY THE JULY 9, 1962, NUCLEAR WEAPONS TEST AS OBSERVED AT COLLEGE, ALASKA

CHARLES R. WILSON

Geophysical Institute, University of Alaska

MASAHISA SUGIURA*

Goddard Space Flight Center

Magnetic effects of the July 9, 1962, high-altitude nuclear weapons test near Johnston Island are studied by means of the record obtained with a rubidium vapor magnetometer at College, Alaska, and the records taken by the U.S. Coast and Geodetic Survey at College and Sitka, Alaska. The first event recorded by the rubidium magnetometer was a sudden increase in the total field 2.1 sec after the detonation; this was followed by a large-amplitude oscillation of period 2.5 to 3 sec, which drove the recorder off scale for several cycles. The first impulse is interpreted as a predominantly longitudinal hydromagnetic wave generated by the detonation. The second event, which commenced 52 sec after the detonation, was recorded by the rubidium magnetometer as a 3 γ negative impulse lasting about 8 sec. The third phase of the bomb effect was a large-amplitude magnetic perturbation of much longer period than the first event, beginning 66 sec after the detonation. This event is interpreted as the arrival of a transverse hydromagnetic wave generated in the magnetosphere at several earth radii and transmitted along the lines of magnetic force to the earth; the transverse wave was generated in the magnetosphere by coupling with the compressional hydromagnetic wave that propagated outward near the equatorial plane from the detonation point.

Magnetic effects of the United States July 9, 1962, high-altitude nuclear weapons test at Johnston Island were observed in Alaska by the U.S. Coast and Geodetic Survey magnetic observatories at College and Sitka with their Ruska rapid-run component magnetometers and by the Geophysical Institute at College with a Varian rubidium vapor total field magnetometer.

The sensitivities and chart speeds of the two instruments at College are: 0.0195 γ/mm and 2032 mm/min (8 inches/min) for the rubidium vapor magnetometer; and for the Ruska component magnetometers 4.6 γ/mm for H, 0.92 min (of arc)/mm for D, 5.0 γ/mm for Z, with a uniform chart speed of 4 mm/min for all three components. Time signals with 1-minute intervals received from the Minitrack Satellite Tracking Station at College were recorded on the rubidium vapor magnetometer record so that the absolute time of sharp magnetic impulses could be determined to within ± 0.06 sec.

Although the magnetic field was moderately disturbed during the night of the test, three distinct groups of hydromagnetic waves were identified as bomb effects at College.

The first event, as recorded by the rubidium vapor magnetometer at College, was a sudden increase in the total field at 0900 11.11 ± 0.06 sec UT; this was followed by a large-amplitude oscillation, of period 2.5 to 3 sec, which drove the recorder off scale for several cycles. Since the fluctuations were too rapid to be resolved in the records of the Coast and Geodetic Survey magnetometers, the polarization of these waves could not be determined. Because of its very large total

*National Academy of Sciences—National Research Council fellow on leave of absence from the Geophysical Institute, College, Alaska.

field variations, this first event is interpreted as the arrival of a predominantly longitudinal hydromagnetic wave generated by the explosion; in general, longitudinal waves are expected to arrive earlier than transverse waves [Dungey, 1958].

The arrival of the second event at 09 h 01 m 01 s±0.5 sec UT was recorded by the rubidium magnetometer as a 3 γ negative impulse lasting about 8 seconds. On the basis of the characteristics of the H, D, and Z rapid-run records, this pulse is interpreted as a predominantly transverse hydromagnetic wave.

Bomke et al. [1960] investigated the magnetic effects from the Argus nuclear bomb tests and proposed that the two signals observed were the longitudinal and transverse hydromagnetic waves generated by the detonations and that these waves were transmitted essentially in ducts concentric with the earth. The travel time for the fast mode along the duct proposed by Bomke et al. from the point directly above Johnston Island to the point directly above College becomes 1.5 sec if values adopted by these authors are used for the height of the duct and the ion density in the duct. The observed travel time was 2.08±0.06 sec. However, there is considerable uncertainty in the ion density at the altitude of the duct, namely 2500 km for the fast mode. Also, the surface distance along the great circle from Johnston Island to College is 5000 km, and the height of the duct is 2500 km; thus it is questionable whether the travel time computed in the above manner for the ducted propagation is meaningful.

The observed travel time can be accounted for at least in order of magnitude as follows: the sudden expansion of the hot plasma created by the detonation generated a longitudinal hydromagnetic wave that subsequently propagated in all directions. Part of the wave energy that was transmitted to a direction obliquely upward was reflected back to the earth's surface by continuous bending downward due to the increasing Alfvén wave velocity with height below the level at which this velocity is a maximum.

Since the wavelength of the longitudinal wave in question is of the same order of magnitude as the dimensions concerned, the straightforward application of ray path theory is probably not valid; hence the propagation path is not analyzed in detail here.

If the second duct for the transverse mode is assumed to be at the height of 1750 km, as in the model proposed by Bomke et al., the travel time for this mode along the duct from the point directly above Johnston Island to the point directly above College is about 8 sec. There was no distinct pulse corresponding to the arrival of this mode.

The third phase of the bomb effect is of considerable interest. This third event was a magnetic perturbation of large amplitude, but the variation was much slower than the first event. Whereas the first event consisted of several oscillations with periods of a few seconds, the time scale involved in the third event was about three minutes. The

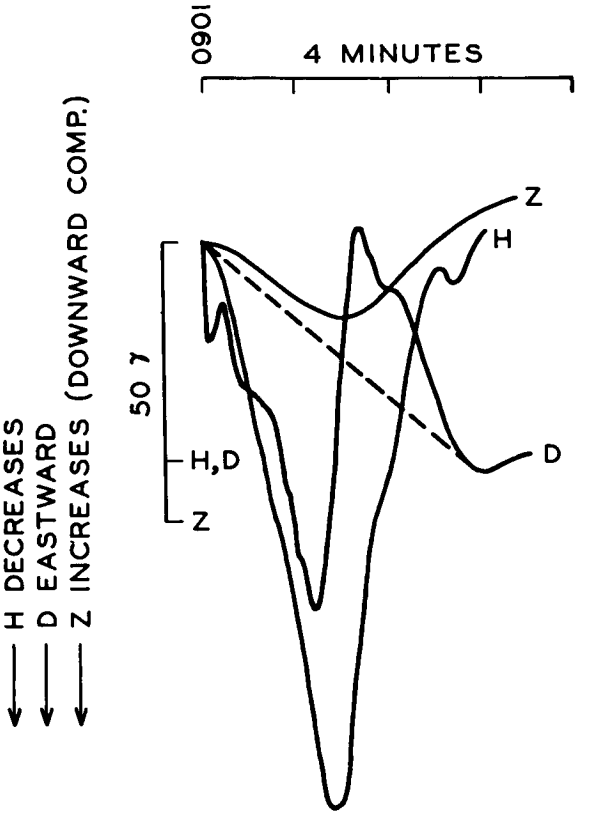


FIGURE 1.—Rapid-run magnetograms of H, D, and Z superimposed to show the phase relationships for elliptically polarized transverse wave at College caused by the July 9, 1962, nuclear bomb test. The dashed line shows the eastward background drift in D.

variations in the three components for the third event are shown in Figure 1.

The horizontal projection of the locus of the end point of the total magnetic perturbation vector is shown in Figure 2. This vector diagram resembles those constructed in a similar manner for the magnetic perturbation at College in the sudden commencement of magnetic storms [Wilson and Sugiura, 1961]. Figure 2 indicates a clockwise rotation of the magnetic perturbation vector.

The variation in declination appears to consist of a quasi-sinusoidal variation and a slow eastward drift. If we assume that this latter change was a slow variation of the field on which the former was superimposed, and if we subtract the slow variation which is represented by broken lines in Figure 1, the vector diagram for the resultant variation becomes more elliptical (Figure 3).

It is of some interest to examine the Z component. If the earth were a perfect conductor, an electromagnetic signal propagating downward from the ionosphere to the earth would be completely reflected at ground so that the Z component of the signal would be zero. Such complete reflection at ground has been assumed by various authors. However, there is some evidence indicating that the ground reflection is not so complete as is usually supposed. (This subject will be discussed in a separate paper.) Even for waves of periods of a few seconds, ground reflection appears to be very incomplete (Dawson and Sugiura, private communication).

Consider an idealized case of an elliptically polarized transverse wave propagating down a field line with a perturbation vector b that varies in magnitude as $b_0 \sin wt$. The arrival of such a wave at College where the dip is 77° would produce the variations in H, D, and Z shown in Figure 4B and a polarization diagram as in

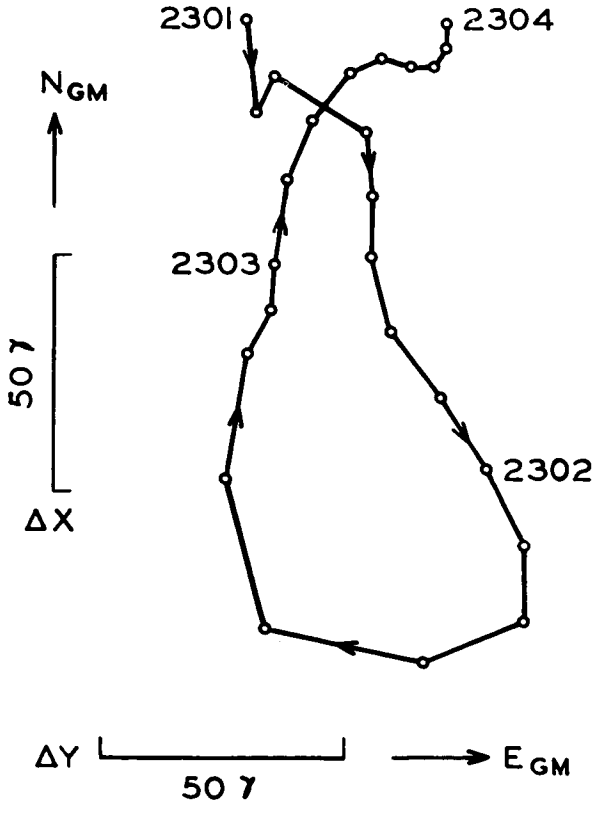


FIGURE 2.—Vector diagram of the transverse wave showing the locus of the end point of the total horizontal disturbance vector ΔH as a function of time.

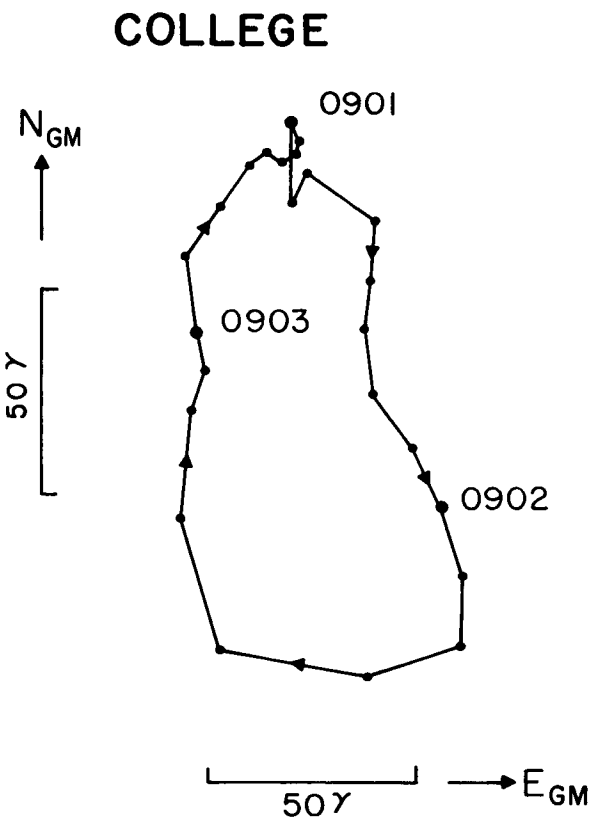


FIGURE 3.—Vector diagram of the transverse wave in figure 2 with the background variation in D subtracted to show the clockwise elliptical polarization more clearly.

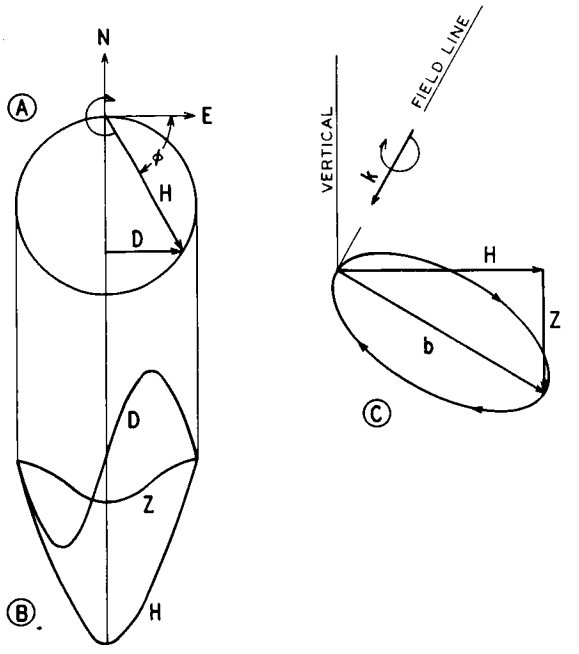


FIGURE 4.—A: Vector diagram for the idealized case in which the disturbance vector b increases as $b_0 \sin\phi$ where $\phi = \omega t$. B: The corresponding variations in H, D, and Z when the dip is 77°. C: Oblique view of the locus of b showing the relation between ΔH and Z. When ΔH is south, Z is downward and in phase with H. K is the propagation vector of the incoming wave.

Figure 4*A*. Referring to the oblique view of the polarization diagram in Figure 4*C*, we can see that when b is directed southward, producing a decrease in H, the change in Z will be an increase. The change in Z with time will be in phase with that of H as shown in Figure 4*B*. The shapes of the H, D, and Z traces in Figure 1 resemble those in Figure 4*B*, which are drawn on the basis that the wave is polarized in the plane perpendicular to the magnetic field line. Thus it appears that the wave observed was indeed an elliptically polarized transverse hydromagnetic wave transmitted along the field line.

We are thus led to speculate that the third event may have been produced in the following way. The longitudinal hydromagnetic wave created by the explosion propagated radially outward away from the earth, and as the wave went out to regions of distances of several earth radii from the earth, it blew out the magnetic field lines sufficiently to generate transverse hydromagnetic waves, which then propagated along the field lines; these transverse waves should bear the characteristics of circular polarization. The circumstance is quite similar to the mechanism which we have proposed for the sudden commencement of a magnetic storm [Wilson and Sugiura, 1961]. In the storm commencement, longitudinal hydromagnetic waves are generated by the encounter of a solar gas cloud with the earth's magnetic field at the outer surface of the magnetosphere and travel toward the earth, whereas in the present case longitudinal hydromagnetic waves were created by the nuclear bomb and propagated outward.

The third phase began about 66 sec after the detonation. We now estimate the transit time of the hydromagnetic perturbation according to our model. We first estimate the time for a hydromagnetic wave to travel from the altitude of the bomb detonation, which is taken to be 400 km, to the point at which the magnetic field line crossing the earth's surface at College intersects the equatorial plane. Since the latitude of Johnston Island is low (19°), we take a path on the equatorial plane out to 5.4 earth radii, and then follow the field line to the earth's surface.

The travel time from 400-km height to 5.4 earth radii was estimated by a numerical integration using the Alfvén wave velocity distribution given by Dessler et al. [1960]; this gives 45 sec.

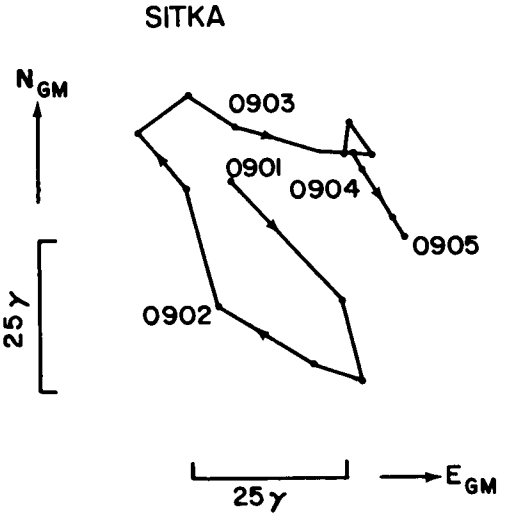


FIGURE 5.—Vector diagram for the transverse wave at Sitka showing clockwise polarization.

According to our observation of the oscillation of the magnetic field lines at College, the period of oscillation for the fundamental mode is about 114 sec. Taking one quarter of this period to be the travel time of a hydromagnetic wave from the equatorial plane at 5.4 earth radii to College along the field line, this travel time becomes 28.5 sec.

Thus the total travel time for the propagation proposed here is about 74 sec. Considering uncertainties in the distribution of Alfvén velocity in the magnetosphere and the crudeness of our estimate, the time of arrival of the third phase can be regarded as being in agreement, in order of magnitude, with the expectation.

If our hypothesis is correct, we might expect to see a similar transverse wave at Sitka where the storm sudden commencement field often shows circular polarization. Figure 5 shows the vector diagram for Sitka, which can be taken to indicate some tendency for the magnetic vector to rotate clockwise.

ACKNOWLEDGMENTS

This project was supported by the Office of Naval Research through contract Nonr 3010(02). We are indebted to the Coast and Geodetic Survey for providing their magnetic records taken at College and Sitka. We wish to thank Dr. R. E. Simpson and Dr. J. P. Heppner for valuable discussions.

REFERENCES

BOMKE, H. A., RAMM, W. J., GOLDBLATT, S., and KLEMAS, V., Global hydromagnetic wave ducts in the exosphere, *Nature*, 185, 1960.

DESSLER, A. J., FRANCIS, W. E., and PARKER, E. N., Geomagnetic storm sudden-commencement rise times, *J. Geophys. Res.*, **65**, 2715, 1960.

DUNGEY, J. W., *Cosmic Electrodynamics*, p. 75, Cambridge University Press, London, 1958.

WILSON, C. R., and SUGIURA, M., Hydromagnetic interpretation of sudden commencements of magnetic storms, *J. Geophys. Res.*, **66**, 4097, 1961.

VI. PLANETOLOGY

1195

ELASTIC MODELS OF THE MANTLE CORRESPONDING TO VARIATIONS IN THE EXTERNAL GRAVITY FIELD

WILLIAM M. KAULA

Goddard Space Flight Center

Static earth-tide theory was modified to include interior loads expressible as spherical harmonics, and elastic moduli were assumed to be functions of radius only. Variations of density from this model and the corresponding stress distributions in the crust and mantle were calculated to correspond to observed variations in the gravitational field plus the surface topography up to 4th-degree spherical harmonics. These solutions were made determinate by imposing the condition of minimization of strain energy. For the elastic parameters derived from seismology, the maximum stress difference obtained from the discrepancy between the observed and equilibrium flattenings of the earth was 163 bars. The maximum stress difference found for the sum of all other terms was 97 bars in the lower mantle and 300 bars in the crust. Displacements were always less than 70 meters. Modifications of the solution which take into account finite strain, creep, and viscous deformation are discussed. A model consisting of a fluid layer 35 to 400 km deep and a rigidity $\frac{1}{2}$ seismic in the rest of the mantle results in a reduction of the maximum stress difference in the mantle to 54 bars and an increase of the maximum displacement to 1500 meters.

INTRODUCTION

Knowledge of the earth's gravity field has been significantly improved in recent years, principally because of the perturbations of close satellite orbits, but also because of the extension of gravimetric coverage and the application of large computers to its analysis. In particular, the long-term perturbations of orbits have yielded very accurate determinations of the low-degree zonal harmonics; the most recent summary and analysis were made by Kozai [1963]. These zonal harmonics are incorporated in the analysis of satellite orbits for tesseral harmonics by Kaula [1963]. The order of magnitude of satellite results has been confirmed by some recent analyses of terrestrial gravimetry: least-squares determination of harmonic coefficients by Uotila [1962] and autocovariance analysis determination of degree variances by Kaula [1959]. The principal features of variations in the earth's gravitational potential indicated by these studies are:

1. Zonal harmonics (normalized) up to degree 4 of the order 10^{-6} times the central term.
2. Zonal harmonics of degrees 5 through 9 of the order 10^{-7}.
3. Tesseral harmonics up to degree 4 of comparable magnitude to the zonal harmonics.
4. Negligible correlation with the corresponding harmonics in the topography for all degrees.

STRESSES IN THE MANTLE

Although there remain some statistically questionable aspects of all these results except the low-degree zonal harmonics, it seems appropriate now to explore their implications with regard to stresses in the mantle. The subject has been discussed generally by O'Keefe [1959], Munk and MacDonald [1960b], and MacDonald [1963]. Licht [1960] has applied the linear viscous model of Vening-Meinesz [Heiskanen and Vening-Meinesz, 1958] to the explanation of the third-degree zonal harmonic. We start at the opposite extreme of an elastic model.

The elastic model is assumed to have elastic moduli which are functions of the radius only. We seek solutions for variations in density which are

functions of latitude and longitude as well, $\delta\rho(r, \Theta, \phi)$, to account for the observed variations in the external gravity field and the surface load constituted by the topography, while at the same time entailing a minimum of shear stress in the elastic model.

The applicable equations in the earth are the equations of equilibrium for a continuous medium:

$$0=\rho(\partial W/\partial x_i)+\sum_j(\partial p_{ij}/\partial x_i) \tag{1}$$

and Poisson's equation:

$$\sum_j(\partial^2 W/\partial x_i^2)=-4\pi G\rho \tag{2}$$

where ρ is density, W is the gravitational potential, x_1, x_2, and x_3 are cartesian coordinates, p_{ij} is the stress tensor, and G is the gravitational constant.

The potential can be considered as consisting of a potential existing in a reference state plus a perturbation:

$$W=W_0+\Psi \tag{3}$$

The corresponding equation for the density is

$$\rho=\rho_0-\int_0^{x_i-x_{0i}}\frac{\partial\rho_0}{\partial x_i}du_i-\rho_0\Delta+\delta\rho \tag{3a}$$

where Δ is the dilation and $\delta\rho$ is a density perturbation from outside the problem: a chemical, structural, or thermal inhomogeneity. In (3a), the rule of summation over the repeated subscript applies, and it will apply in all subsequent equations.

We make the assumptions that: (1) the reference state is one of fluid equilibrium, (2) the stress-strain relationship is elastic, and (3) the displacements $u_i=x_i-x_{0i}$ are small enough so that only linear terms must be considered.

These assumptions enable us to write the constitutive equation as

$$p_{ij}=\delta_{ij}\lambda\Delta+\mu[(\partial u_i/\partial x_i)+(\partial u_j/\partial x_i)] \tag{4}$$

$$\lambda=k-2\mu/3, \tag{4a}$$

where δ_{ij} is the Kronecker delta, μ is the modulus of rigidity, λ is the Lamé constant, and k is the bulk modulus.

The problem thus far is identical with that of earth tides, which has been treated by Takeuchi [1950], Molodenskii [1953], and Alterman et al. [1959], with one exception: the term $\delta\rho$ in the expression for the density, (3a). Substituting the expressions from (3) and (3a) into (1) and (2) and neglecting the products of small quantities, we obtain terms additional to those of the earth tide problem of

$$\delta\rho\partial W_0/\partial x_i \tag{1a}$$

on the right of (1) and

$$-4\pi G\ \delta\rho \tag{2a}$$

on the right of (2).

If the hydrostatic reference state is subtracted out, and the equations converted to spherical coordinates, as in equations 7 through 10 of Alterman et al. [1959], the density perturbation will appear only in the radial equation of equilibrium (7) as

$$-\delta\rho g_0 \tag{1b}$$

where g_0 is the negative of the radial gradient of the reference potential. Following Alterman et al. [1959] further, we express the displacement as a spheroidal vector spherical harmonic:

$$\mathbf{u}=U(r)\begin{Bmatrix}0\\0\\S_n(\theta,\varphi)\end{Bmatrix}+V(r)\begin{Bmatrix}\partial S_n(\theta,\varphi)/\partial\theta\\[\partial S_n(\theta,\varphi)/\partial\varphi]/\sin\theta\\0\end{Bmatrix} \tag{5}$$

where $S_n(\theta, \varphi)$ is a surface spherical harmonic, and we express the perturbation of the potential as

$$\psi=P(r)\,S_n(\theta,\varphi) \tag{6}$$

Furthermore, we express $\delta\rho$ as

$$\delta\rho=D(r)\,S_n(\theta,\varphi) \tag{7}$$

Substituting (5) and (6) into equations 7 through 10 of Alterman et al. [1959], and substituting (7) for our extra terms (1b) and (2a) we get the additional term $-Dg_0$ on the left side of their equation 23 and the additional term $-4\pi GD$ on the right of their equation 25.

Finally, following their conversion of variables, we get the six first-order equations:

$$y_1' = -\frac{2\lambda y_1}{(\lambda+2\mu)r} + \frac{y_2}{(\lambda+2\mu)} + \frac{\lambda n(n+1)}{(\lambda+2\mu)r} y_3$$

$$y_2' = \left[-4\rho_0 g_0 r + \frac{4\mu(3\lambda+2\mu)}{(\lambda+2\mu)}\right]\frac{y_1}{r^2} - \frac{4\mu}{(\lambda+2\mu)r} y_2$$

$$+\left[n(n+1)\rho_0 g_0 r - \frac{2\mu(3\lambda+2\mu)n(n+1)}{(\lambda+2\mu)}\right]\frac{y_3}{r^2}$$

$$+\frac{n(n+1)}{r} y_4 - \rho_0 y_5 + D g_0$$

$$y_3' = -\frac{y_1}{r} + \frac{y_3}{r} + \frac{y_4}{\mu}$$

$$y_4' = \left[g_0\rho_0 r - \frac{2\mu(3\lambda+2\mu)}{(\lambda+2\mu)}\right]\frac{y_1}{r^2} - \frac{\lambda}{(\lambda+2\mu)}\frac{y_2}{r}$$

$$+\frac{2\mu}{(\lambda+2\mu)}[\lambda(2n^2+2n-1)$$

$$+2\mu(n^2+n-1)](y_3/r^2) - (3y_4/r) - (\rho_0 y_5/r)$$

$$y_5' = 4\pi G\rho_0 y_1 + y_6$$

$$y_6' = -\frac{4\pi G\rho_0 n(n+1)}{r} y_3 + \frac{n(n+1)}{r^2} y_5 - \frac{2y_6}{r}$$

$$-4\pi G D \quad (8)$$

where the y_i's are the radial factors, respectively, of the radial displacement, the radial stress, the tangential displacement, the tangential stress, the potential perturbation, and the gradient of the potential perturbation less the radial displacement contribution thereto. Primes denote derivatives with respect to r.

We abbreviate (8) as

$$y_i' = Q_{ij} y_j + w_i D \quad (9)$$

At least one Q_{ij} is of the order of $1/r$ for all j, so all y_j's must be zero at the origin. Furthermore, some Q_{i1}, Q_{i3}, and Q_{i5} are $0(1/r^2)$, so $y_1'(0)$, $y_3'(0)$, and $y_5'(0)$ must all be zero, leaving three constants of integration, $y_2'(0)$, $y_4'(0)$, and $y_6'(0)$, for a solid core. For a fluid core, there are only two independent equations (cf. Longman [1963]):

$$y_5' = y_6 . \quad (10)$$

$$y_6' = \{[n(n+1)/r^2] - (4\pi\rho^2/\lambda)\} y_5 - (2/r) y_6$$

For a fluid core with radius c, we took as constants of integration $y_6'(0)$, $y_1(c)$, and $y_3(c)$. The conditions at the surface $r=a$ with a layer of surface density σ_n are

$$y_2(a) = -g_0(a)\sigma_n$$

$$y_4(a) = 0$$

$$[(n+1)/a] y_5(a) + y_6(a) = 4\pi G\sigma_n \quad (11)$$

If we consider the external gravitational field as known, with coefficient C_n for the spherical harmonic component of the potential, there is a fourth surface condition,

$$y_5(a) = C_n \quad (11')$$

If it is also assumed that $D(r)$ is known throughout, the problem is overdetermined. The simplest assumption is that $D(r)$ is known except for a constant multiplier κ, which becomes a fourth unknown in addition to the three constants of integration. If σ_n is zero and C_n is the exterior potential due to $D(r)$ in a perfectly rigid earth, the load Love number k_n'' for the internal load $D(r)$ comparable to the k_n' calculated for surface loads by Takeuchi et al. [1962], and Longman [1963] is

$$k_n'' = (1-\kappa)/\kappa \quad (12)$$

κ, like k_n'', is thus a measure of the response of the earth to the internal load $\delta\rho$; the exterior potential perturbation C_n will be a combination of the potential due to $\delta\rho$ itself and that due to the deformation of the earth by the load $\delta\rho$.

If $D(r)$ is assumed to involve two or more unknown parameters, further assumptions are required in order to make the problem determinate. The most logical assumption is that the shear strain energy is minimized; i.e.,

$$\int \mu(r) e_{ij}' e_{ij}' \, dV = \text{minimum} \quad (13)$$

where μ is the rigidity, the integration is over the volume of the earth, and e_{ij}' is the deviatoric

strain tensor [Jeffreys, 1959 p. 12]

$$e_{ij}' = e_{ij} - \tfrac{1}{3}\delta_{ij} e_{kk} \tag{14}$$

To connect e_{ij} with the y_i's, we use the equations for the strain in spherical coordinates [Love, 1927, p. 56], (applying a factor $\frac{1}{2}$ to the off-diagonal components of strain to be consistent with tensor convention) and eliminate dy_1/dr and dy_3/dr by (8)

$$e_{\theta\theta} = 2(S_{nm}/r)y_1 + (2/r)(\partial^2 S_{nm}/\partial\theta^2)y_3$$

$$e_{\varphi\varphi} = 2\frac{S_{nm}}{r}y_1 + \frac{2}{r\sin\theta}\left(\frac{1}{\sin\theta}\frac{\partial^2 S_{nm}}{\partial\varphi^2} + \cos\theta\frac{\partial S_{nm}}{\partial\theta}\right)y_3$$

$$e_{rr} = -\frac{4\lambda}{\lambda+2\mu}\frac{S_{nm}}{r}y_1 + \frac{2n(n+1)\lambda}{(\lambda+2\mu)}\frac{S_{nm}}{r}y_3 + \frac{2S_{nm}}{\lambda+2\mu}y_2$$

$$e_{\theta\varphi} = \frac{2}{r\sin\theta}\left(\frac{\partial^2 S_{nm}}{\partial\theta\partial\varphi} - \cot\theta\frac{\partial S_{nm}}{\partial\varphi}\right)y_2$$

$$e_{\varphi r} = (\mu\sin\theta)^{-1}(\partial S_{nm}/\partial\varphi)y_4$$

$$e_{r\theta} = \mu^{-1}(\partial S_{nm}/\partial\theta)y_4 \tag{15}$$

We express (15) in matrix form

$$e(\theta, \varphi, r) = \mathbf{M}(\theta, \varphi)\mathbf{y}(r) \tag{16}$$

The shear strain energy per unit volume can then be expressed as

$$s(\theta, \varphi, r) = u(r)\mathbf{y}^{\tau}\mathbf{M}^{\tau}\mathbf{J}\mathbf{M}\mathbf{y} \tag{17}$$

where the τ superscript denotes the transpose and

$$\mathbf{J} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} & 0 & 0 & 0 \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & 0 & 0 & 0 \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

The Runge–Kutta or other numerical method of solution of (9) (see, e.g., Kopal [1955]) can be adapted to express the variables y_i at level k in terms of the y_i at level $k-1$ and the density function $D(r)$ at levels $k-1$, $k-\frac{1}{2}$, and k:

$$\mathbf{y}_k = \mathbf{T}_k\mathbf{y}_{k-1} + \mathbf{w}_k D_k + \mathbf{w}_{2k}D_{k-1/2} + \mathbf{w}_{3k}D_{k-1} \tag{18}$$

If the density function $D(r)$ is expressed as a function of a few parameters $\mathbf{d}$, (18) can be simplified to the form

$$\mathbf{y}_k = \mathbf{T}_k y_{k-1} + \mathbf{W}_k\mathbf{d} \tag{19}$$

By repeated substitution of the $k-1$, $k-2$, etc., equations into (19) we obtain

$$\mathbf{y}_k = \mathbf{U}_k\mathbf{c} + \mathbf{V}_k\mathbf{d} \tag{20}$$

where $\mathbf{c}$ is a vector of the constants of integration, $y_6'(0)$, $y_1(c)$, and $y_3(c)$,

$$U_k = \prod_{j=1}^{k}\mathbf{T}_j$$

and

$$V_k = \sum_{i=2}^{k}\prod_{j=i}^{k}\mathbf{T}_j\mathbf{W}_{j-1} + \mathbf{W}_k$$

and $\mathbf{T}_1$ is appropriately modified to fit the initial conditions.

The surface conditions (equations 11 and 11′) for m levels of integration are expressible as

$$\mathbf{a} = \mathbf{C}\mathbf{y}_m = \mathbf{C}\mathbf{U}_m\mathbf{c} + \mathbf{C}\mathbf{V}_m\mathbf{d} \tag{21}$$

The system of equations (17), (20), and (21), along with the condition expressed by (13), is mathematically identical to that of generalized least squares (see, e.g., Arley and Buch [1950, pp. 196–198]) with the vector of corrections to observations

$$\mathbf{x} = (\mathbf{c}\vdots\mathbf{d})^{\tau} \tag{22}$$

condition equation coefficient matrix

$$\mathbf{F} = (\mathbf{C}\mathbf{U}_m\vdots\mathbf{C}\mathbf{V}_m) \tag{23}$$

and covariance matrix

$$\mathbf{W} = \sum_k \mu(r_k)(\mathbf{U}_k\vdots\mathbf{V}_k)^{\tau}\left(\int\mathbf{M}^{\tau}\mathbf{J}\mathbf{M}\,d_\sigma\right)(\mathbf{U}_k\cdots\mathbf{V}_k) \tag{24}$$

where the integration is over the surface of the sphere. The solution of the system of equations (21) through (24), subject to (13), by the method of Lagrangian multipliers is

$$\mathbf{x} = \mathbf{W}\mathbf{F}^{\tau}(\mathbf{F}\mathbf{W}\mathbf{F}^{\tau})^{-1}\mathbf{a} \tag{25}$$

TABLE 1.—*Mantle Model Corresponding to the External Gravitational Field Minimum Shear Strain Energy, Gutenberg Model of Elastic Parameters*

Coefficient	External potential, earth units	Topography surface density, equivalent earth units	Crustal density anomaly, g/cm³	Maximum mantle density anomaly, g/cm³	Maximum mantle stress difference various radii		Total shear strain energy,* ergs
					Δp, bars	R, earth radii	
$\Delta\bar{C}_{20}$	5.0×10^{-6}	4.65×10^{-6}	0.012	5.7×10^{-4}	163	0.56	22.4×10^{22}
$\bar{C}_{22}$	1.84	−3.9	0.021	−1.2	70	0.56	4.6
$\bar{S}_{22}$	−1.71	−0.34	−0.002	−1.3	65	0.56	2.4
$\bar{C}_{30}$	0.98	−2.50	0.014	−1.4	32	0,56	2.4
$\bar{C}_{31}$	1.77	−1.52	0.012	−2.2	60	0.56	5.4
$\bar{S}_{31}$	−0.11	1.16	−0.006	1.2	5	0.87	0.1
$\bar{C}_{32}$	0.34	−4.45	0.022	−2.9	15	0.56	2.3
$\bar{S}_{32}$	0.08	3.94	−0.018	2.7	3	0.56	1.6
$\bar{C}_{33}$	−0.31	0.72	−0.004	0.3	11	0.56	0.2
$\bar{S}_{33}$	0.74	4.47	−0.019	3.3	55	0.56	2.8
$\bar{C}_{40}$	−0.41	2.68	−0.014	2.1	18	0.78	0.6
$\bar{C}_{41}$	−0.21	−1.68	0.007	−1.2	8	0.87	0.4
$\bar{S}_{41}$	0.46	−2.46	0.013	−1.8	19	0.77	1.2
$\bar{C}_{42}$	−0.03	−4.0	0.018	−2.9	9	0.98	1.6
$\bar{S}_{42}$	0.32	0.6	−0.002	0.5	12	0.56	0.1
$\bar{C}_{43}$	0.50	3.0	−0.012	2.2	15	0.86	1.4
$\bar{S}_{43}$	0.16	−1.8	0.009	−1.3	7	0.98	0.4
$\bar{C}_{44}$	−0.24	−0.19	0.000	−0.2	9	0.56	0.2
$\bar{S}_{44}$	0.55	4.25	−0.018	3.1	20	0.56	2.5

* Neglecting interactions between different harmonics.

Solutions were made for density distributions corresponding to spherical harmonics of the external gravitational field up to the harmonic $Y_{4,4}$ given by Kaula [1963] and the Gutenberg model density and elasticity parameters as given by Takeuchi et al. [1959]. The surface layer coefficients were derived from the harmonic analysis of the topography by G. J. Bruins, as described by Vening-Meinesz [1959]. In these solutions the density parameters for the mantle and the crust were kept separate. Two different types of parameters **d** were used to represent the density anomalies in the mantle: layers and polynomial coefficients. About the same answers were obtained by the two methods, but those from the polynomial coefficients varied more smoothly. In general, no appreciable reduction in the summation of strain energy was obtained by using more than four parameters to represent the density variations for each harmonic in the mantle. One parameter was used to represent the density anomaly in the crust.

As a test of the program, tidal and surface-load Love numbers were also calculated by fixing $D(r) = 0$ and appropriately modifying the surface conditions. The answers agreed more closely with those of Longman [1963] than with those of Takeuchi et al. [1962].

The results for the polynomial density-variation solution are summarized in Table 1. The potential and surface density coefficients are in units such that the radius of the earth, the mass of the earth, and the gravitational constant are all unity, and all coefficients apply to normalized spherical harmonics $\bar{S}_{nm}$ so that the integral of $\bar{S}_{nm}^2$ over the unit sphere is 4π. The $\Delta\bar{C}_{20}$ is the discrepancy between the observed value and that corresponding to fluid equilibrium [O'Keefe, 1959]. The

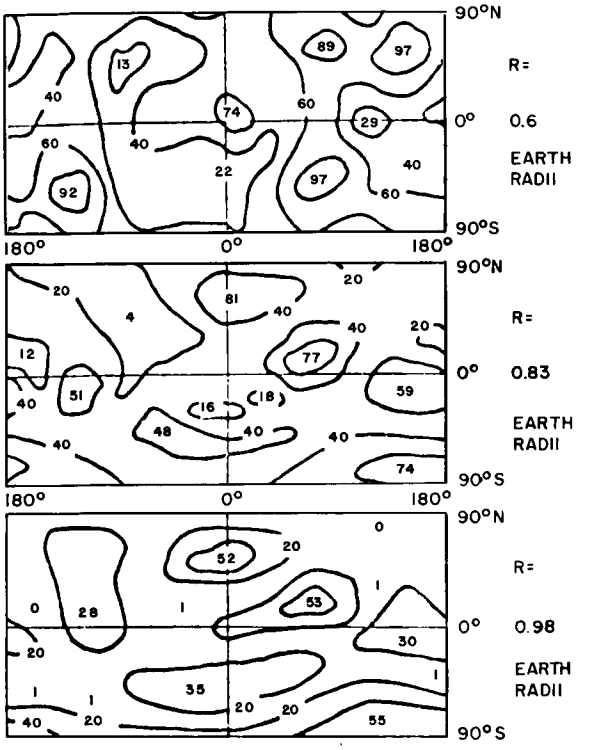


FIGURE 1.—Principal stress differences in bars, rigidity and bulk modulus assumed from seismology (Gutenberg model).

maximum shear stress obtained for $\Delta\bar{C}_{20}$ is consistent with the 100 bars strength estimated by Munk and MacDonald [1960a, p. 280].

Figures 1 and 2 sum up the effects of all coefficients except $\Delta\bar{C}_{20}$ in the form of maps of the maximum stress difference and the radial component of displacement at three selected levels within the mantle. The displacements have, in general, a negative correlation with the variations in the external field [Kaula, 1963, figure 1]. The stress differences, which are quadratic functions of the displacement gradients, do not show such a clear pattern, although maximums generally occur in areas where the correlation between the topography and the gravitational field is most negative. The appreciable variability of the stress differences suggests that a significantly different solution might be obtained by applying a yield stress limit. Also calculated were strain energy densities; the maximum shear strain energy density found in the mantle was 1100 ergs/cm^3.

As might be expected under the criterion of strain energy minimization, for most terms the density anomaly in the crust accomplished almost complete isostatic compensation of the surface topography. However, in satisfying the condition of the known external gravitational field the compensation is never exact: generally, for terms in which the surface layer has a sign opposite to that of the gravitational field term, there is an overcompensation, and if it has the same sign, there is an undercompensation. The maximum density anomaly in the mantle nearly always fell immediately below the crust.

A manifestly oversimplified physical model has two possible values: to give a notion of the quantities involved and to suggest how a more realistic (and hence probably more complicated) model might be developed. The stresses and energies found for the elastic model are probably a crude approximation to those in the actual earth, but the deformations obviously fall far short of explaining what occurs. We therefore want to

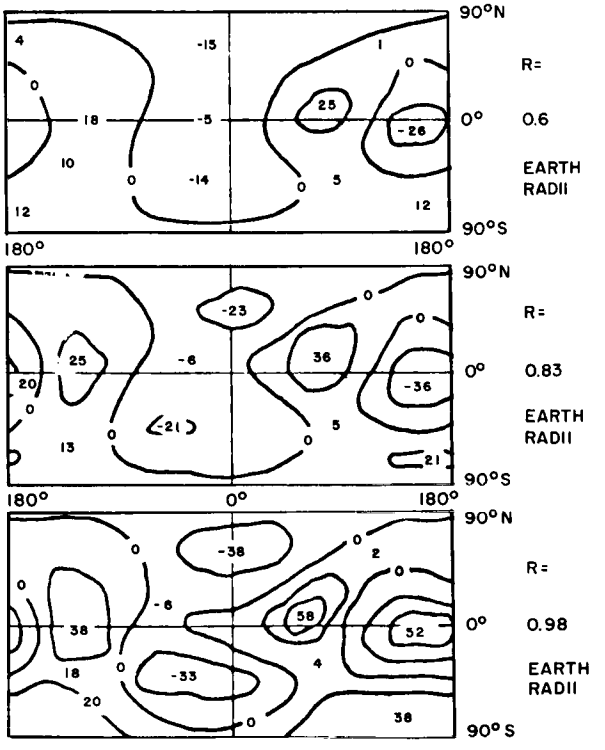


FIGURE 2.—Radial displacements in meters, rigidity and bulk modulus assumed from seismology (Gutenberg model).

examine in turn the possible modifications of the model in the categories of (1) finite strain elasticity: the strain is constant, but the stress is significantly affected by higher-order terms than those in equation 4; (2) creep: the strain rate is appreciable, but the stress remains a function of only the strain; and (3) viscous deformation: the stress is a function of the strain rate.

FINITE STRAIN ELASTICITY

Finite strain is significant within the earth in two respects: first, the density and elastic moduli for small superimposed loads, such as seismic waves, must vary with pressure and temperature in a way consistent with actual materials; second, the higher-order terms may contribute appreciably to the stress: i.e., products of displacement gradients may be significant in calculating strain, and products of strain components may be significant in calculating stress.

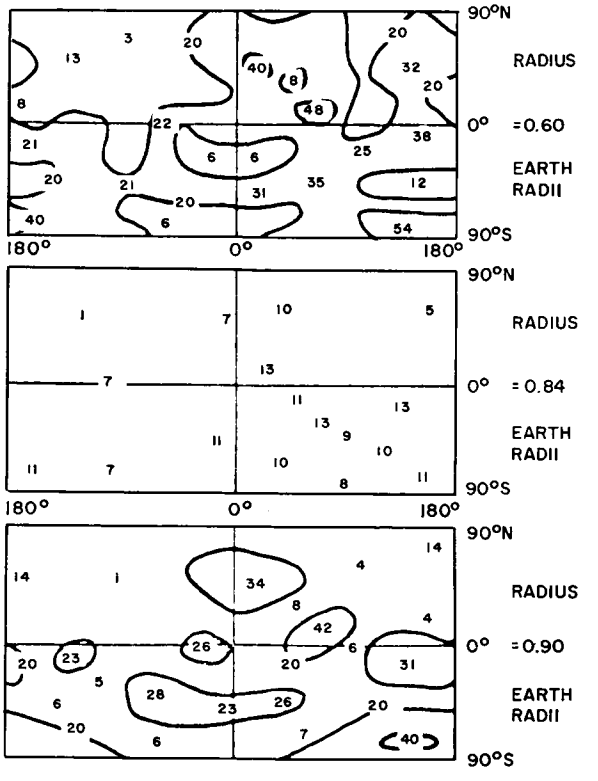


FIGURE 3.—Principal stress differences in bars, bulk modulus assumed from seismology throughout fluid layer 35 to 400 km deep; and ½ seismic rigidity 400 to 2900 km deep.

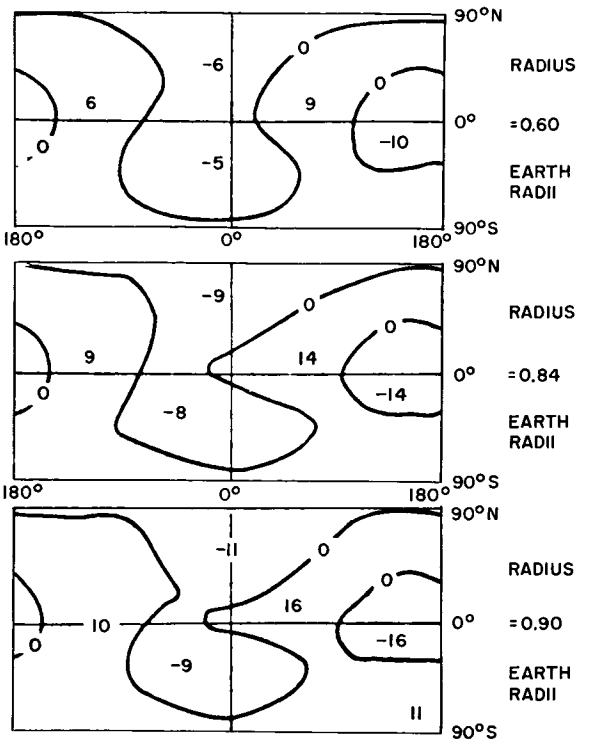


FIGURE 4.—Radial displacements in meters, bulk modulus assumed from seismology throughout; fluid layer 35 to 400 km deep; and ½ seismic rigidity 400 to 2900 km deep.

The first aspect of the finite strain theory has been applied most extensively by Birch [1952, 1961] to the problem of deducing composition or phase changes in the mantle from seismic velocities. It is of concern in our present problem only to the extent that the density and elasticity parameters assumed should be materially plausible.

The second aspect of finite strain theory seems of dubious applicability to the earth's mantle, since the strains involved in the minimum energy described above are less than 10^{-4}. If higher-order terms were taken into account, the initial state of stress would no longer be simply additive; hence the solution according to strain energy minimization or other criterion would be a function of the initial state of stress. This problem was investigated by Jeffreys [1943], who assumed an implicit nonlinear constitutive equation in the form of a stress matrix with one higher-order term of the form $x_i x_k S_n$. The model was a

homogeneous shell with density layers at the upper and lower boundaries. The radial factors of the displacements were expressed in terms of four functions of the form $A_s+B_sr^2$. The eight parameters A_s and B_s were adjusted to minimize the strain energy or stress difference while satisfying the boundary conditions separately for each harmonic. The reduction in stress obtained was moderate, even though interactions between different harmonics were neglected and it was not made clear whether the constitutive equation implied by the solution was realistic.

In our formulation of the problem, it seems very probable that considerations of nonelasticity are of much greater concern than nonlinear elasticity terms.

CREEP

There have been several recent discussions of the geophysical application of experimental evidence and theoretical models of creep: Orowan [1960], Griggs and Handin [1960], Jeffreys [1958, 1959], Jeffreys and Crampin [1960], MacDonald [1961], Lomnitz [1962], Weertman [1962], Scheidegger [1963], Stacey [1963], and Donath and Faill [1963]. There appear to be appreciable differences of terminology, interpretation, and opinion in these discussions, so we attempt to summarize the principal conclusions in order to decide how our mathematical model should be modified.

1. Transient creep, or elastic afterworking. Consideration of the phase lags and dampings in the response of the earth to periodic disturbances leads to creep models of the form [Jeffreys, 1958; Jeffreys and Crampin, 1960; MacDonald, 1961; Lomnitz, 1962]

$$\mathbf{e}(t)=\mathbf{p}\mu^{-1}[1+\psi(t)] \tag{26}$$

for the strain $\mathbf{e}$ at a time $t>0$ due to a constant stress $\mathbf{p}$ applied at $t=0$. "Creep functions" $\Psi(t)$ which appear to fit phenomena having periods up to the 430 days of the free nutation are

$$\psi(t)=q\log(1+at) \tag{27}$$

and

$$\psi(t)=[(1+at)^{\alpha}-1]q/\alpha \tag{28}$$

The time scale of these phenomena falls far short of that in our problem, but Jeffreys and Crampin [1960] find that their numerical values in (28) would permit an appreciable part of a second-degree harmonic to survive for more than 10^9 years. MacDonald [1961] objects to the rules (27) and (28) because they imply an infinite population of relaxation times in a linear superposition model and because they do not lead to frequency independence of dissipation, as observed for frequencies above 1 cps, but these considerations have no apparent bearing on the long term problem.

2. Delayed elasticity. Glasses at low temperature can be regarded as elastic, but their full response to loads is appreciably delayed. If the irregularities of grain boundaries, etc., are such that the material of the mantle can be regarded as amorphous, there may be a significant amount of delayed elasticity. Orowan [1960] suggests that delayed elasticity may account for the postglacial rise. In this case, elasticity would have to prevail to a considerable depth in the mantle, rather than only in the crust, as considered by Vening-Meinesz [Heiskanen and Vening-Meinesz, 1958].

3. Steady-state creep, or elastoviscosity. Crystalline substances at low shear stresses above a "creep strength" deform very slowly. This steady-state creep can usually be expressed as [Weertman, 1962; Stacey, 1963]

$$\dot{\mathbf{e}}=C(p)\exp(-Q/kT) \tag{29}$$

Where $\dot{\mathbf{e}}$ is the strain rate, Q is an activation energy, k is the Boltzmann constant, T is the absolute temperature, and C is an exponential or, near melting, a power of the shear stress $\mathbf{p}$. Jeffreys [1958] considers steady-state creep to be of no importance in the earth; Stacey [1963] suggests that it is the dominant rheological mode in the mantle.

4. Uniform flow. If the shear stress in a crystalline substance exceeds a certain yield point, rapid plastic deformation or uniform flow takes place. At confining pressures above 1000 bars, the zone of failure widens, but cohesion is retained after failure [Griggs and Handin, 1960; Donath and Faill, 1963].

Instability of creep in metals has been observed to result in the rapid propagation of intense shear

bands, or faulting without fracture [Orowan, 1960]. This "shear melting" is suggestive of a mechanism for earthquakes or magma formation.

Estimates of the creep strength or yield point at the temperatures and pressures prevailing in the mantle seem to be rather speculative. Orowan [1960] states that the creep strengths of crystalline materials close to the melting point are of the order of 100 or 10 bars, or even less. Stacey [1963], on the other hand, obtains creep strengths (defined as the "stress at which creep reaches [a] just observable value") in excess of 500 bars, mainly, it seems, from the observation of Scheidegger [1963, p. 159] that departures from isostasy which result in stresses smaller than about 4000 bars can exist "almost indefinitely." Weertman [1962] does not give a creep strength, but he does say that a steady-state creep law of the form of (29) breaks down at stresses above 100 to 1000 bars, which presumably defines the yield point. Griggs and Handin [1960] make no sharp distinction between creep and uniform flow but suggest that shear melting could occur in the mantle at shear stresses as low as 100 bars.

We seem to have authority for a wide range of hypotheses as to creep strengths and yield points in the mantle. In particular, in comparing the stress differences in figure 1 with the values suggested for the strengths, it appears entirely possible that the mantle can be assumed to be in a state in which creep is taking place, but in which stress is a function of the strain, not of the strain rate. Such a state would occur if the density irregularities $\delta\rho$ were not removed by the creep in the 10^8 or 10^9 years time scale of our problem. Taking the situation at any instant $\tau=0$ as a reference state, we find that the rigidity μ in (4) would be replaced by a pseudo-rigidity

$$\mu_p(t)=\mu/[1+\psi(t)] \qquad (30)$$

TABLE 2.—*Mantle Model Corresponding to the External Gravitational Field Minimum Shear Strain Energy, Fluid Layer 35 to 400 km Deep, ½ Gutenberg Model Rigidity in Lower Mantle*

Coefficient	External potential, earth units	Topography surface density, equivalent earth units	Crustal density anomaly, g/cm³	Maximum mantle density anomaly, g/cm³	Maximum mantle stress difference at various radii		Total shear strain energy,* ergs
					Δp, bars	R, earth radii	
$\Delta\bar{C}_{20}$	5.0×10^{-6}	4.65×10^{-6}	0.0146	-2.0×10^{-4}	31	0.56	1.84×10^{23}
$\bar{C}_{22}$	1.84	−3.9	0.0019	−1.0	13	0.56	0.19
$\bar{S}_{22}$	−1.71	−0.34	−0.0042	0.7	12	0.56	0.19
$\bar{C}_{30}$	0.98	−2.50	0.0034	−3.5	22	0.937	0.40
$\bar{C}_{31}$	1.77	−1.52	0.0080	−6.1	35	0.937	1.39
$\bar{S}_{31}$	−0.11	1.16	0.0002	0.4	2	0.937	0.05
$\bar{C}_{32}$	0.34	−4.45	0.0010	−1.4	7	0.937	0.06
$\bar{S}_{32}$	0.08	3.94	0.0028	0.03	1	0.56	0.03
$\bar{C}_{33}$	−0.31	0.72	−0.0011	1.1	6	0.937	0.04
$\bar{S}_{33}$	0.74	4.47	0.0065	−2.3	13	0.937	0.32
$\bar{C}_{40}$	−0.41	2.68	−0.0013	2.7	16	0.937	0.14
$\bar{C}_{41}$	−0.21	−1.68	−0.0024	1.2	7	0.937	0.05
$\bar{S}_{41}$	0.46	−2.46	0.0017	−3.0	16	0.937	0.08
$\bar{C}_{42}$	−0.03	−4.0	−0.0025	−0.09	5	0.82	0.03
$\bar{S}_{42}$	0.32	0.6	0.0025	−1.9	9	0.937	0.10
$\bar{C}_{43}$	0.50	3.0	0.0052	−2.9	13	0.937	0.26
$\bar{S}_{43}$	0.16	−1.8	0.0006	−1.1	5	0.937	0.02
$\bar{C}_{44}$	−0.24	−0.19	−0.0018	1.5	7	0.937	0.05
$\bar{S}_{44}$	0.55	4.25	0.0062	−3.1	16	0.937	0.34

* Neglecting interactions between different harmonics.

where

$$\psi(t) = \int_0^t A(t-\tau)\mathbf{p}(\tau)\, d\tau > 0$$

which is formally similar to the transient creep function of MacDonald [1961]. Equation 30 makes $\mu_p(t)$ a function of latitude and longitude, which in turn would not permit development of the solution in the form of (8). To see how the stresses and displacements might be modified in an extreme case, however, a solution of the system of equations 21 to 24 subject to (13) was made for a model in which $\mu_p(t)$ was zero in the low-velocity zone 35 to 400 km deep, and equal to $\frac{1}{2}$ the Gutenberg model rigidity throughout the rest of the mantle. This solution is summarized in Table 2. The principal result was that the incompressibility of the fluid layer, in addition to that of the core, reduced the shear stresses required by supporting much of the load. Specifically, there was (1) an increase in the radial displacements of the crust by a factor of about 30, (2) a decrease in the shear stresses in the crust to about $\frac{1}{4}$, (3) a reduction of the stresses in the mantle to about $\frac{1}{5}$ for the second degree, $\frac{1}{2}$ for the third degree, and $\frac{4}{5}$ for the fourth degree, (4) a concentration of stresses in the mantle near the solid-liquid boundaries, and (5) a density-anomaly maximum at the top of the elastic part of the mantle. The reduction of the $\Delta\bar{C}_{20}$ stress to a value comparable to that for $\bar{C}_{31}$ suggests that reduction of rigidity similar to that of the hypothesized model could take place in the 10^7 years indicated by the lag in adjustment of the rotational bulge. The relatively small reduction in stresses for the fourth-degree terms suggests that higher-degree variations could not be reasonably supported by such a model, and it is consistent with the indications from satellite zonal harmonics and autocovariance analysis of gravimetry that there is an appreciable drop-off in the magnitude of variations of the gravitational field the fourth degree.

VISCOUS DEFORMATION

We may ask two questions: Is large-scale convective motion now taking place in the mantle? If convection is taking place, is the stress which supports the density irregularities primarily a function of the strain rate—i.e., is it viscous (Newtonian or otherwise)? Answers to the first question appear to depend on the background of the answerer, the two extremes being the geothermists [Lubimova, 1960; MacDonald, 1963], who consider that convection must be limited in time and place (of the nature of shear melting) in order that the observed heat flow not be greatly exceeded, and the paleomagnetists [Runcorn, 1962; Stacey, 1963], who need to explain evidence of continental drift of the order of 3 cm/year. Answers to the second question are even vaguer because of the lack of experimental data, the uncertainty as to the appropriate rheological theory, and the lack of an adequate mathematical solution for convection with significant variation of parameters and a finite yield stress, as discussed by MacDonald [1963]. If it is assumed that stress is a function of strain rate d_{ij} with constitutive equation $p_{ij} = \rho\, d_{kk}\, \delta_{ij} + 2\eta\, d_{ij}$, where η is the viscosity, a solution for the displacement rate field analogous to (4) to (25) for the displacement field would be somewhat more complicated, since the external gravitational field would still be a function of the displacement field. We should expect, however, that the order of magnitude of the rates $\mathbf{v}$ could be estimated from the displacements $\mathbf{u}$ by

$$\mathbf{v} \approx \mathbf{u}\mu/\eta \qquad (31)$$

The pseudo-viscosity η most commonly quoted is 10^{22} poises, deduced by Vening-Meinesz [Heiskanen and Vening-Meinesz, 1958, pp. 365–370] and others from a decay time of about 5300 years and a linear dimension of about 1200 km for postglacial uplift. The Newtonian viscous model requires that the decay time vary inversely as the diameter of the load; however, Crittenden [1963] recently obtained a decay time of about 4000 years for the 180-km-diameter Bonneville Lake area uplift, or a pseudo-viscosity η of 10^{21} poises.

Taking $\mu = 10^{12}$, $\mathbf{u} = 3 \times 10^3$, and $\eta = 10^{22}$ cgs in (31), we obtain velocities of the order of 10 cm/year. Licht [1960] objected to such a high rate because it required improbably high efficiency of heat transport from core to surface; however, his discussion seems to be based on the implicit assumption that the radiogenic heat in

the mantle is less than 10^{-2} of the 1.6 ergs/g/yr estimated for chondritic composition [MacDonald, 1959]. Runcorn [1962] also appears to neglect radiogenic heating in the mantle in showing that convective velocities of the order of 10 cm/yr are obtained with temperature differences of only 0.2 degree centigrade, thus yielding a surface heat flow much less than observed. A more fundamental weakness of mantle-wide convection models is the assumption that the pseudo-viscosity of 10^{22} poises can be applied to organized convection with a characteristic length of 2900 km; Crittenden's result emphasizes the danger of such an extrapolation.

To reconcile the gravitational and thermal evidence with a Newtonian viscous model appears to require either a mantle-wide system with viscosity coefficient in excess of 10^{23} poises or a system with a characteristic length of 180 km or less and a viscosity of 10^{21} poises or less. The first model leaves unexplained many surface evidences of large-scale motion; the second model has not been worked out. Therefore, further development of convective models seems worth while, as well as critical re-examination of the evidences of continental drift [Munk and MacDonald, 1960a, pp. 251–262, 282–285] and study of the problem of distribution of radiogenic heating.

CONCLUSIONS

The indications we have found as to the state of stress in the earth's interior from the low-degree, or long-wave, variations in the external gravitational field are minimal in two respects. First, the condition of strain energy minimization has been imposed; the actual density distribution could conceivably be quite different. Second, superimposed on the low-degree variations may be higher-degree variations of sufficient magnitude to increase appreciably the stress differences above those here calculated. If the scale of significant change in the earth is appreciably smaller than the wave-lengths of the low-degree harmonics—as suggested by the narrowness of the belts of surface manifestation of tectonic activity and the relatively shallow origin of seismic activity—then the long-wave variations in the gravity field supply evidence of only the passive background for the processes currently important.

The development of models more realistic than the elastic cannot be purely mechanical because of the rheological uncertainties. To provide some limitation to the possible solutions, we should include energy flows and distribution of heat sources in the problem. The desirability of incorporating the thermal aspects, as well as shorter-wave variations in the gravitational field, suggests a more statistical approach to keep the problem manageable. The inputs would be the spectrums of gravity, topography, heat flow, etc., variations in the form of degree variances

$$\sigma_n^2 = \sum_{m=0}^{n} (\bar{C}_{nm}^2 + \bar{S}_{nm}^2)$$

together with some measure of the correlation between the different observed quantities. The solutions sought would be the spectrums of density, displacement, displacement rate, heat sources, etc., at various levels within the earth under specified conditions such as the strain energy minimization.

ACKNOWLEDGMENT

I am grateful to W. D. Putney for doing much of the programming for the computations described in this paper.

REFERENCES

ALTERMAN, Z., JAROSCH, H., and PEKERIS, C. L., Oscillations of the earth, *Proc. Roy. Soc. London*, **A252,** 80–95, 1959.

ARLEY, N., and BUCH, K. R., *Introduction to the Theory of Probability and Statistics*, 236 pp., John Wiley & Sons, New York, 1950.

BIRCH, F., Elasticity and constitution of the earth's interior, *J. Geophys. Res.*, **57,** 227–286, 1952.

———, Composition of the earth's mantle, *Geophys. J.*, **4,** 295–310, 1961.

CRITTENDEN, M. D., Jr., Effective viscosity of the earth derived from isostatic loading of Pleistocene Lake Bonneville, *J. Geophys. Res.*, **68**(19): 1963.

DONATH, F. A., and FAILL, R. T., Ductile faulting in experimentally deformed rocks (abstract), *Trans. Am. Geophys. Union*, **44,** 103–104, 1963.

GRIGGS, D., and HANDIN, J., Observations on fracture and a hypothesis of earthquakes, *Rock Deformation, Geol. Soc. Am. Mem.*, **79,** 347–364, 1960.

HEISKANEN, W. A., and VENING-MEINESZ, F. A., *The Earth and Its Gravity Field*, 470 pp., McGraw-Hill Book Company, New York, 1958.

JEFFREYS, H., The stress-differences in the earth's shell, *Monthly Notices Roy. Astron. Soc., Suppl.* **5,** 71–85, 1943.

———, A modification of Lomnitz's law of creep in rocks, *Geophys. J.*, **1,** 92–95, 1958.

———, *The Earth*, 4th ed., 420 pp., Cambridge University Press, 1959.

———, and CRAMPIN, S., Rock creep: a correction, *Monthly Notices Roy. Astron. Soc.*, **121,** 571–577, 1960.

KAULA, W. M., Statistical and harmonic analysis of gravity, *J. Geophys. Res.*, **64,** 2401–2421, 1959.

———, Tesseral harmonics of the gravitational field and geodetic datum shifts derived from camera observations of satellites, *J. Geophys. Res.*, **68,** 473–484, 1963.

KOPAL, Z., *Numerical Analysis*, 556 pp., John Wiley & Sons, New York, 1955.

KOZAI, Y., The potential of the earth derived from satellite motions, *Proc. Intern. Union Theoret. and Appl. Mech. Symp. Dynamics of Satellites, Paris*, Springer-Verlag, Berlin, 1963.

LICHT, A. L., Convection currents in the earth's mantle, *J. Geophys. Res.*, **65,** 349–353, 1960.

LOMNITZ, C. A., Application of the logarithmic creep law to stress wave attenuation in the solid earth, *J. Geophys. Res.*, **67,** 365–368, 1962.

LONGMAN, I. M., A Green's function for determining the deformation of the earth under surface mass loads, 2, Computations and numerical results, *J. Geophys. Res.*, **68,** 485–496, 1963.

LOVE, A. E. H., *A Treatise on the Mathematical Theory of Elasticity*, 4th ed., 643 pp., Cambridge University Press, London, 1927; republished by Dover Publications, New York, 1944.

LUBIMOVA, H. A., On conditions of magmatism, origin and role of volcanic activity in the thermal regime of earth's crust, *J. Phys. Earth, Tokyo*, **8,** 17–27, 1960.

MACDONALD, G. J. F., Calculations on the thermal history of the earth, *J. Geophys. Res.*, **64,** 1967–2000, 1959.

———, Geophysical implications of satellite geodesy, *Proc. Intern. Symp. Use of Artificial Satellites for Geodesy, Washington*, North-Holland Publishing Company, Amsterdam, in press, 1963.

MACDONALD, J. R., Theory and application of a superposition model of internal friction and creep, *J. Appl. Phys.*, **32,** 2385–2398, 1961.

MOLODENSKII, M. S., Elastic tides, free nutation, and some questions of the structure of the earth (in Russian), *Tr. Geofiz. Inst., Akad. Nauk SSSR*, **19**(146), 1–42, 1953.

MUNK, W. H., and MACDONALD, G. J. F., *The Rotation of the Earth*, 323 pp., Cambridge University Press, New York, 1960(*a*).

———, and ———, Continentality and the gravitational field of the earth, *J. Geophys. Res.*, **65,** 2169–2172, 1960(*b*).

O'KEEFE, J. A., Zonal harmonics of the earth's gravitational field and the basic hypothesis of geodesy, *J. Geophys. Res.*, **64,** 2389–2399, 1959.

OROWAN, E., Mechanism of seismic faulting, *Rock Deformation, Geol. Soc. Am. Mem.*, **79,** 323–346, 1960.

RUNCORN, S. K., Towards a theory of continental drift, *Nature*, **193,** 311–314, 1962.

SCHEIDEGGER, A. E., *Principles of Geodynamics*, 2nd ed., 362 pp., Academic Press, New York, 1963.

STACEY, F. D., The theory of creep in rocks and the problem of convection in the earth's mantle, *Icarus*, **1,** 304–313, 1963.

TAKEUCHI, H., On the earth tide of the compressible earth of variable density and elasticity, *Trans. Am. Geophys. Union*, **31,** 651–689, 1950.

———, PRESS, F., and KOBAYASHI, N., Rayleigh-wave evidence for the low-velocity zone in the mantle, *Bull. Seismol. Soc. Am.*, **49,** 355–364, 1959.

———, SAITO, M., and KOBAYASHI, N., Statical deformations and free oscillations of a model earth, *J. Geophys. Res.*, **67,** 1141–1154, 1962.

UOTILA, U. A., Harmonic analysis of world-wide gravity material, *Publ. Isostasy Inst. Intern. Assoc. Geodesy*, **39,** 17 pp., Helsinki, 1962.

VENING-MEINESZ, F. A., The results of the development of the earth's topography in spherical harmonics up to the 31st order; provisional conclusions, *Konikl. Ned. Akad. Wetenschap. Proc.*, **B62,** 115–125, 1959.

WEERTMAN, J., Mechanism for continental drift, *J. Geophys. Res.*, **67,** 1133–1140, 1962.

TIDAL DISSIPATION IN THE MOON

WILLIAM M. KAULA

Goddard Space Flight Center

Dissipation of tidal energy in the moon was calculated under the assumption that it can be represented as due to imperfect elasticity. If the factor $1/Q$ for dissipation per cycle is assumed to be 1/100 for distortional strain energy and 1/1000 for dilatational strain energy, the heat now being generated in the moon by tides is less than 0.010 erg/g/yr.; i.e., negligible compared to radioactive heating by a chondritic composition. Tidal heating would be compared to radioactive heating, however, if the semimajor axis of the moon's orbit was one-third as great, so the mechanism does limit the possible history of the moon's orbit. Also, appreciable tidal dissipation would cause thermal stresses and be conducive to convection, since it is much greater in the center than near the surface of the moon, and has a non-uniform distribution in latitude (and longitude as well, if rotation is synchronous with revolution).

Urey et al. [1959] suggested that tidal dissipation may have significantly contributed to heating of the moon's interior in the past. Kopal [1963] has calculated the tidal dissipation in a moon assumed to be a fluid with Newtonian viscosity. However, since tidal distortion is a relatively high frequency phenomenon, it seems a more reasonable extrapolation from experience to assume that tidal dissipation in the moon occurs as a result of imperfect elasticity with a factor $1/Q$ for dissipation per cycle comparable to those estimated for the earth's mantle from polar tides, free oscillations, and latitude variation.

The tidal disturbing function W is

$$W = (GM^*/r^*) \sum_{l=2}^{\infty} (r/r^*)^l P_l \ (\cos S) \qquad (1)$$

Where P_l is a Legendre polynomial; r, ϕ, λ are radius, latitude, and longitude in a moon-fixed coordinate system; the asterisked quantities refer to the disturbing body; GM^* is the product of the gravitational constant and the mass; and S is the arc from (ϕ^*, λ^*) to (ϕ, λ). We apply the addition theorem to (1):

$$W = \frac{GM^*}{r^*} \sum_{l=2}^{\infty} \left(\frac{r}{r^*}\right)^l \sum_{m=0}^{l} \frac{(l-m)!}{(l+m)!} (2-\delta_{0m}) \cdot P_{lm} \ (\sin\phi) \, P_{lm} \ (\sin\phi^*) \cdot [\cos m\lambda \cos m\lambda^* + \sin m\lambda \sin m\lambda^*] \qquad (2)$$

Where P_{lm} is the Legendre associated function and δ_{0m} is the Kronecker delta.

To obtain the variation in time of r^*, ϕ^*, λ^*, we must express them in terms of the earth's orbit referred to the moon. For the present orbit this is most simply done by using the numerical values of coefficients as given by Brown's theory of lunar motion, as recently described by Harrison [1963]. Because we wish to investigate the effects of changing the orbital parameters, and because the necessary computer subroutines had already been written for another purpose (to a degree of detail superfluous to the present problem), we used a different development [Kaula, 1961] based on the assumption that the orbit can be considered a Keplerian ellipse at a fixed inclination to the moon's equator with secularly moving node and perigee. For the tidal problem, the most significant omissions under this assumption are short-period perturbations of the semimajor axis and the longitude by the sun. For the present orbit, the largest of these terms (those containing h in the arguments of equations 6 to 8 of Harrison [1963]) have a ratio of about 0.2 to the terms arising from the ellipticity of the orbit. This ratio would vary directly with variation in the semimajor axis of the lunar orbit but would stay about the same with variation in the inclination or eccentricity. It is consistent, then, with the unavoidable crudeness of

our estimates of the dissipation factors $1/Q$ to assume a purely elliptic orbit.

We apply the transformation of equations 7 to 28 in Kaula [1961] to $r^{*l-l}P_{lm}\ (\sin\phi^*)[\cos m\lambda^*, \sin m\lambda^*]$ in (2):

$$W=\frac{GM^*}{a^*}\sum_{l=2}^{r}\left(\frac{r}{a^*}\right)^l\sum_{m=0}^{l}\frac{(l-m)!}{(l+m)!}(2-\delta_{0m})P_{lm}\ (\sin\phi)$$
$$\cdot\sum_{p,q}F_{lmp}(i^*)G_{lpq}(e^*)\left[\cos m\lambda\begin{Bmatrix}\cos\\ \sin\end{Bmatrix}_{l-m\ \mathrm{odd}}^{l-m\ \mathrm{even}}\right.$$
$$\{(l-2p)\omega^*+(l-2p+q)M^*+m(\Omega^*-\theta)\}$$
$$+\sin m\lambda\begin{Bmatrix}\sin\\ -\cos\end{Bmatrix}_{l-m\ \mathrm{odd}}^{l-m\ \mathrm{even}}\{(l-2p)\omega^*$$
$$\left.+(l-2p+q)M^*+m(\Omega^*-\theta)\}\right] \qquad (3)$$

where a^*, e^*, i^*, Ω^*, ω^*, and M^* are the Keplerian elements of the earth's orbit referred to the moon's equator and a ceparture point thereon fixed with respect to inertial space; $F_{lmp}(i^*)$ and $G_{lpq}(e^*)$ are polynomials of the sine and cosine of the inclination and of the eccentricity, respectively, and θ is the 'lunar sidereal time': the angle between the inertially fixed departure point and the point on the moon from which selenographic longitudes are measured. We assume the rate $\dot\theta$ to be constant, which is equivalent to neglecting the physical libration.

We abbreviate (3) as

$$W=\sum_{l=2}^{\infty}r^l\sum_{m=0}^{l}P_{lm}\ (\sin\phi)$$
$$\cdot\sum_{g}a_{lmg}\left[\cos m\lambda\begin{Bmatrix}\cos\\ \sin\end{Bmatrix}_{l-m\ \mathrm{odd}}^{l-m\ \mathrm{even}}(\sigma_{lmg}t-t_{lmg})\right.$$
$$\left.+\sin m\lambda\begin{Bmatrix}\sin\\ -\cos\end{Bmatrix}_{l-m\ \mathrm{odd}}^{l-m\ \mathrm{even}}(\sigma_{lmg}t-t_{lmg})\right] \qquad (4)$$

where the single subscript g replaces the subscript pair p and q; the amplitude

$$a_{lmg}=\frac{GM^*}{a^*}\left(\frac{r}{a^*}\right)^l\frac{(l-m)!}{(l-m)!}(2-\delta_{0m})F_{lmp}(i^*)G_{lpq}(e^*) \qquad (5)$$

and the rate

$$\sigma_{lmg}=(l-2p)\dot\omega^*+(l-2p+q)\dot M^*+m(\dot\Omega^*-\dot\theta) \qquad (6)$$

The strain energy per unit volume, divided into shear (or distortional) and compressive (or dilatational) terms [Jeffreys, 1959, p. 12] is

$$E=\mu e_{ij}'e_{ij}'+(k/6)e_{mm}e_{nn} \qquad (7)$$

where summation is taken over repeated subscripts. In (7), μ is the rigidity, k is the bulk modulus, e_{ij} is the strain tensor and e_{ij}' is the part of the strain tensor expressing departures from symmetry:

$$e_{ij}'=e_{ij}-\tfrac{1}{3}\delta_{ij}e_{kk} \qquad (8)$$

where δ_{ij} is the Kronecker delta.

If the tidal disturbing function is expressed as a sum of spherical harmonics,

$$W=\sum_{l=2}^{\infty}r^l\sum_{m=0}^{l}P_{lm}\ (\sin\phi)$$
$$\cdot[A_{lm}(t)\ \cos m\lambda+B_{lm}(t)\ \sin m\lambda] \qquad (9)$$

the strain tensor e_{ij} at any point $(r,\ \phi,\ \lambda)$ can be expressed as

$$e_{ij}=\sum_{l,m}[A_{lm}(t)\epsilon_{ijlmc}(r,\phi,\lambda)+B_{lm}(t)\epsilon_{ijlms}(r,\phi,\lambda)] \qquad (10)$$

where $\epsilon_{ijlmc,s}$ is the response of a planetary model of specified shear modulus, bulk modulus, and density to a unit coefficient in the tidal disturbing function.

Comparing (4) and (9), we have

$$A_{lm}(t)=\sum_{g}a_{lmg}\begin{Bmatrix}\cos\\ \sin\end{Bmatrix}_{l-m\ \mathrm{odd}}^{l-m\ \mathrm{even}}(\sigma_{lmg}t-t_{lmg}) \qquad (11)$$

$$B_{lm}(t)=\sum_{g}a_{lmg}\begin{Bmatrix}\sin\\ -\cos\end{Bmatrix}_{l-m\ \mathrm{odd}}^{l-m\ \mathrm{even}}(\sigma_{lmg}t-t_{lmg}) \qquad (12)$$

Substituting (10), (11), and (12), into (7)

and sorting out the algebra which results yields

$$E=\sum_{l,m,g,u,v,h}\Big[(S_{1lmguvh}+C_{1lmguvh})\begin{Bmatrix}\cos\\ \sin\end{Bmatrix}_{b\,\mathrm{odd}}^{b\,\mathrm{even}}\{(\sigma_{lmg}+\sigma_{uvh})t-(t_{lmg}+t_{uvh})\}+(S_{2lmguvh}+C_{2lmguvh})$$

$$\cdot\begin{Bmatrix}\sin\\ \cos\end{Bmatrix}_{b\,\mathrm{odd}}^{b\,\mathrm{even}}\{(\sigma_{lmg}+\sigma_{uvh})t-(t_{lmg}+t_{uvh})\}+(S_{3lmguvh}+C_{3lmguvh})\begin{Bmatrix}\cos\\ \sin\end{Bmatrix}_{b\,\mathrm{odd}}^{b\,\mathrm{even}}\{(\sigma_{lmg}-\sigma_{uvh})t-(t_{lmg}-t_{uvh})\}$$

$$+(S_{4lmguvh}+C_{4lmguvh})\begin{Bmatrix}-\sin\\ \cos\end{Bmatrix}_{b\,\mathrm{odd}}^{b\,\mathrm{even}}\{(\sigma_{lmg}-\sigma_{uvh})t-(t_{lmg}-t_{uvh})\}\Big] \quad (13)$$

where $b=l+u-m-v$; (u, v) summations start at (l, m); and

$$S_{1lmguvh}=a_{lmg}a_{uvh}(\mu/2)\,(\epsilon_{ijlmc}'\epsilon_{ijuvc}'-\epsilon_{ijlms}'\epsilon_{ijuvs}')\,(2-\delta_{uv}{}^{lm})$$

$$C_{1lmguvh}=a_{lmg}a_{uvh}(k/12)\,(\epsilon_{iilmc}\epsilon_{jjuvc}-\epsilon_{iilms}\epsilon_{jjuvs})\,(2-\delta_{uv}{}^{lm})$$

$$S_{2lmguvh}=a_{lmg}a_{uvh}(\mu/2)\,(\epsilon_{ijlmc}'\epsilon_{ijuvs}'+\epsilon_{ijlms}'\epsilon_{ijuvc}')\,(2-\delta_{uv}{}^{lm})\tfrac{1}{2}[(-1)^{u-v}+(-1)^{l-m}]$$

$$C_{2lmguvh}=a_{lmg}a_{uvh}(k/12)\,(\epsilon_{iilmc}\epsilon_{jjuvs}+\epsilon_{iilms}\epsilon_{jjuvc})\,(2-\delta_{uv}{}^{lm})\tfrac{1}{2}[(-1)^{u-v}+(-1)^{l-m}]$$

$$S_{3lmguvh}=a_{lmg}a_{uvh}(\mu/2)\,(\epsilon_{ijlmc}'\epsilon_{ijuvc}'+\epsilon_{ijlms}'\epsilon_{ijuvs}')\,(2-\delta_{uv}{}^{lm})\,(-1)^{u-v}$$

$$C_{3lmguvh}=a_{lmg}a_{uvh}(k/12)\,(\epsilon_{iilmc}\epsilon_{jjuvc}+\epsilon_{iilms}\epsilon_{jjuvs})\,(2-\delta_{uv}{}^{lm})\,(-1)^{u-v}$$

$$S_{4lmguvh}=a_{lmg}a_{uvh}(\mu/2)\,(\epsilon_{ijlmc}'\epsilon_{ijuvs}'-\epsilon_{ijlms}'\epsilon_{ijuvs}')\,(2-\delta_{uv}{}^{lm})\tfrac{1}{2}[1-(-1)^{b}]$$

$$C_{4lmguvh}=a_{lmg}a_{uvh}(k/12)\,(\epsilon_{iilmc}\epsilon_{jjuvs}-\epsilon_{iilms}\epsilon_{jjuvc})\,(2-\delta_{uv}{}^{lm})\tfrac{1}{2}[1-(-1)^{b}] \quad (14)$$

For any particular term of subscripts l, m, u, v, g, and h, the energy dissipated in one cycle of duration $2\pi/(\sigma_{lmg}+\sigma_{uvh})$ will be

$$\Delta E_{lmguvh+}=2\pi[\,|\,S_{1lmguvh}/Q_s+C_{1lmguvh}/Q_c\,|+|\,S_{2lmguvh}/Q_s+C_{2lmguvh}/Q_c\,|] \quad (15)$$

and in one cycle of duration $2\pi/(\sigma_{lmg}-\sigma_{uvh})$

$$\Delta E_{lmguvh-}=2\pi[\,|\,S_{3lmguvh}/Q_s+C_{3lmguvh}/Q_c\,|+|\,S_{4lmguvh}/Q_s+C_{4lmguvh}/Q_c\,|] \quad (16)$$

To obtain the contributions to energy dissipation per unit time, we multiply (15) and (16) by the absolute values of the rates per unit time, $|\,\sigma_{lmg}+\sigma_{uvh}\,|$ and $|\,\sigma_{lmg}-\sigma_{uvh}\,|$. We sum over all subscripts to obtain the mean dissipation rate at a particular point fixed in the moon:

$$\dot{E}_d(r,\,\phi,\,\lambda)=\sum_{l,m,g,u,v,h}[\Delta E_{lmguvh+}\,|\,\sigma_{lmg}+\sigma_{uvh}\,|+\Delta E_{lmguvh-}\,|\,\sigma_{lmg}-\sigma_{uvh}\,|] \quad (17)$$

If the rotation synchronous with revolution, as it is at present,

$$\dot{\omega}^*+\dot{M}^*+\dot{\Omega}^*-\dot{\theta}=0 \quad (18)$$

there are terms that will contribute only through their amplitude a_{lmg} but not through their rate σ_{lmg}. In the synchronous case, a reference longitude must be fixed. If this reference longitude is the mean direction of the earth, all terms containing $\sin\{m(\omega^*+M^*+\Omega^*-\theta)\}$ are zero and all terms containing $\cos\{m(\omega^*+M^*+\Omega^*-\theta)\}$ are unity; i.e., in (10) the contribution to $A_{lm}(t)$ will be a_{lmg} for $l-m$ even and 0 for $l-m$ odd, and to $B_{lm}(t)$ it will be 0 for $l-m$ even and $-a_{lmg}$ for $l-m$ odd. Practically the only term from which this effect is significant is $(l, m, p, q)=(2, 2, 0, 0)$.

Also, there will be degeneracies for $m=0$ in all

cases and for $m \neq 0$ in the synchronous case, requiring the combination of terms before proceeding as in equations 13 to 16. In these cases, the rate for the term of subscripts (l, m, p, q) will be the negative of the rate for the term of subscript $(l, m, l-m-p, -q)$. If the rate is taken as that of the (l, m, p, q) term, the amplitude for the cosine coefficient will be

$$a_{lmg} = a_{lmh} + (-1)^{l-m} a_{lmi} \tag{19}$$

and for the sine coefficient

$$b_{lmg} = a_{lmh} - (-1)^{l-m} a_{lmi} \tag{20}$$

where the subscript correspondence is h with (p, q) and i with $(l-m-p, -q)$, and a_{lmh} and a_{lmi} are computed by (5). Then (14) must be modified so that b_{lmg} and b_{uvh} coefficients appear in front of the ϵ_{ijlms}, ϵ_{ijuvs}, etc., terms in place of a_{lmg} and a_{uvh}.

Another set of degeneracies occurring in the synchronous case arises because terms of subscript (l, m, p, q) will have rates equal to terms of subscript $(l, m+2i, p-i, q)$, where i is any integer.

Including terms for which the disturbing function rate is zero in effect makes the energy dissipation rate a function of the constant value of the strain, which raises the question of whether strains from other than tidal causes should also be considered. Since we are interested in the dissipation over geological durations of time, these terms should perhaps be omitted because in such time we would expect nonoscillating strains to be removed by an elastic processes. But the dissipation rate obtained would then be an absolute minimum for the assumed Q. The moments of inertia of the moon indicate that it now contains strains larger than tidal, so leaving in the nonoscillating tidal terms should yield a dissipation rate unlikely to be too high.

The quantities a_{lmg}, σ_{lmg}, $l=2$, and $m=0$, 1, 2 were calculated from (5) and (6), for a variety of lunar orbits, using subroutines for $F_{lmp}(i^*)$ and $G_{lpq}(e^*)$ originally devised for analysis of close satellite orbits, and computing the rates $\dot{\omega}^*$, $\dot{M}^*$, and Ω^* by the methods described by Kaula [1961].

The strain tensors ϵ_{ijlm}' and ϵ_{iilm} were calculated using the formulation of the earth-tide problem of Alterman et al. [1959], which has also been used by Takeuchi et al. [1962] and Longman [1963]. In this formulation, the basic variables are the radial factors of vector spherical harmonic expressions of the displacements, stresses, and potential terms: y_1, of the radial displacement; y_2, of the compressive stress; y_3, of the tangential displacement; y_4, of the shear stress; y_5, of the potential; and y_6, of the potential gradient less the contribution thereto of the radial displacement. The equations of equilibrium then become a system of six-first-order equations:

$$dy_i/dr = P_{ij} y_j \tag{21}$$

The P_{ij}'s are functions of r, k, μ, g, and ρ. Three columns of P_{ij} have terms of $O(r^{-2})$, so that the requirement of regularity at the origin eliminates three constants of integration. The three surface conditions of zero tangential stress, zero radial stress, and the potential gradient being related to the potential as a spherical harmonic in free space make the problem determinate. After solving (21) numerically for the y_i's corresponding to a particular harmonic $Y_{lm} = r^l S_{lm}$, the contribution to the strain matrix ϵ_{ij} is calculated by (here θ is colatitude, ϕ is longitude, and λ is $k - \frac{2}{3}\mu$):

$$\epsilon_{\theta\theta} = 2(S_{lm}/r) y_1 + (2/r)(\partial^2 S_{lm}/\partial\theta^2) y_3$$

$$\epsilon_{\phi\phi} = 2(S_{lm}/r) y_1 + (2/r \sin\theta) \cdot \left(\frac{1}{\sin\theta} \frac{\partial^2 S_{lm}}{\partial\phi^2} + \cos\theta \frac{\partial S_{lm}}{\partial\theta} \right) y_3$$

$$\epsilon_{rr} = -\frac{4\lambda}{\lambda+2\mu} \cdot \frac{S_{lm}}{r} y_1 + \frac{2l(l+1)\lambda}{\lambda+2\mu} \cdot \frac{S_{lm}}{r} y_3 + \frac{2 S_{lm}}{\lambda+2\mu} y_2$$

$$\epsilon_{\theta\phi} = \frac{2}{r \sin\theta} \left(\frac{\partial^2 S_{lm}}{\partial\theta\partial\phi} - \cot\theta \frac{\partial S_{lm}}{\partial\phi} \right) y_2$$

$$\epsilon_{\phi r} = (\mu \sin\theta)^{-1} (\partial S_{lm}/\partial\phi) y_4$$

$$\epsilon_{r\theta} = \mu^{-1} (\partial S_{lm}/\partial\theta) y_4 \tag{22}$$

Equation 22 can be derived using the equations of Love [1937, p. 56], applying a factor of $\frac{1}{2}$ to the off diagonal components to be consistent

with tensor convention. The ϵ_{ijlm}'s from the Y_{lm}'s were used in (14).

The pole tide suggests a Q of more than 100 for the earth at a 14-month period [Munk and MacDonald, 1960, p. 162], while the latitude variations indicate a Q of about 40 at the same period [Munk and MacDonald, 1960, p. 148; Jeffreys, 1959, pp. 255–259]. The free oscillations of the earth indicate Q's of 170 to 400 for periods less than an hour [Benioff et al., 1961; Ness et al., 1961], except for a Q of 7500 for the S_0^0, the only mode which is purely compressive [Ness et al., 1961]. Models for rock creep proposed by Jeffreys and Crampin [1960], J. R. MacDonald 19[61], and Lomnitz [1962] suggest Q's between 40 and 100 for semimonthly and monthly periods. A shear Q_s of 100 thus seems a reasonable compromise. Considering that excitation from the atmosphere and oceans may maintain the S_0^0 free oscillation [Ness et al., 1961], we assume a compressive Q_c of 1000.

The strain energies were calculated for several lunar models proposed by Harrison [1963]. However, since the uncertainty in Q reduces this problem to one of estimating order of magnitude, this discussion will be limited to a homogeneous moon of density 3.34 g cm^{-3}, rigidity 7.38×10^{11} dyne cm^{-2}, and bulk modulus 1.23×10^{12} dyne cm^{-2}. The Love numbers obtained for this model

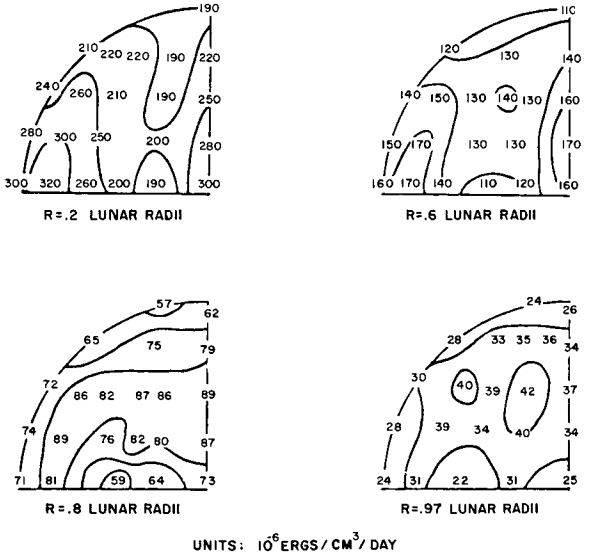


FIGURE 1.—Present tidal energy dissipation in the moon, assuming a shear Q of 100 and a compressive Q of 1000.

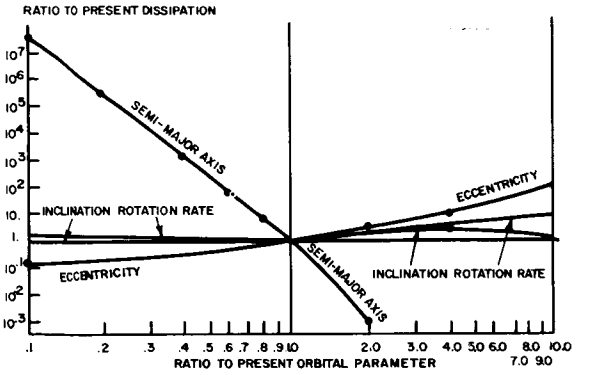


FIGURE 2.—Variation of tidal energy dissipation in the moon with one-at-a-time variation in orbital parameters.

by the numerical solution of (15) were 0.0344 for h, 0.0195 for k, and 0.0095 for l.

The thermal histories of lunar models with chondritic composition have been calculated by G. J. F. MacDonald [1959]. Even with a cold origin, these models come very close to melting at depths in the moon exceeding 500 km. The chondritic composition used had radioactive contents of 8.0×10^{-4}, 1.1×10^{-8}, and 4.4×10^{-8} g/g for potassium, urnaium, and thorium, respectively, which yield a thermal energy output of 1.59 ergs/g/year at present and 12.8 ergs/g/year 4.5×10^9 years ago. Hence for tidal dissipation to be significant, it should contribute of the order of 5 ergs/g/year, or 0.04 erg/cm^3/day.

The results for the present orbit of the moon are shown in figure 1 in the form of maps of the energy dissipation at four levels within the moon. Since the dissipation is symmetric about the equator and two meridians at right angles, we need to show only one octant for each level. The evident features are, first, that the tidal dissipation is at present a negligible source of heat and, second, that the distribution of the heating is extremely nonuniform both radially and laterally. This variability of distribution suggests that, if the moon's orbit had ever been such that the over-all heating was appreciable, large thermal stresses would have resulted, leading to convection or some other form of mass motion.

The amplitudes and rates were therefore calculated for different orbital specifications. Each element was varied in turn, the others being

held fixed at the present values: (1) the semimajor axis (holding the rotation synchronous with revolution about the earth); (2) the eccentricity; (3) the inclination; and (4) the rotation rate. The results are displayed in figure 2, in the form of curves showing variation in the average ratio to the present dissipation with variation in the orbital elements. The variation is particularly marked with variation in the semimajor axis. If the semimajor axis were only one-third as great as it is now, the criterion of 0.04 erg/cm^3/day would be exceeded for most of the moon. At the secular acceleration calculated by Munk and MacDonald [1960], the moon would have been at this distance about 10^9 years ago.

The pattern of energy dissipation shown in Figure 1 is composed of even-degree harmonics symmetric about the equator: (4, 4), (4, 2), (4, 0), (2, 2), and (2, 0). In a moon close enough so that heating by tidal dissipation was large enough to cause convection, the second-degree terms in its mass distribution would be determined mainly by the gravitational attraction of the earth. However, the fourth-degree terms would be determined by the convective pattern. It will therefore be interesting, when variations in the gravitational field and external form of the moon are better determined, to find out whether these fourth-degree terms are markedly larger than other terms, such as the third-degree terms. If they are, it will be a strong indication that the moon was once close enough for heating by tidal dissipation to cause convection, and hence it would lend further evidence as to the moon's origin.

In conclusion we can say that heating by tidal dissipation is currently insignificant (unless Q factors estimated from the earth's mantle are wrong by a factor of about 100) but that this study confirms the suggestion of Urey et al. [1959] that it would have been important in the past if the moon were much closer to the earth.

ACKNOWLEDGMENT

W. D. Putney contributed greatly to this work by programming most of the elastic deformation calculations.

REFERENCES

ALTERMAN, Z., JAROSCH, H., and PEKERIS, C. L., Oscillations of the earth, *Proc. Roy. Soc. London*, **252,** 80–95, 1959.

BENIOFF, H., PRESS, F., and SMITH, S., Excitation of the free oscillations of the earth by earthquakes, *J. Geophys. Res.*, **66,** 605–620, 1961.

HARRISON, J. C., An analysis of the lunar tides, *J. Geophys. Res.*, **68**(14): 1963.

JEFFREYS, H., *The Earth*, 4th ed., 420 pp., Cambridge University Press, 1959.

———, and CRAMPIN, S., Rock creep: a correction, *Monthly Notices Roy. Astron. Soc.*, **121,** 571–577, 1960.

KAULA, W. M., Analysis of gravitational and geometric aspects of geodetic utilization of satellites, *Geophys. J.*, **5,** 104–133, 1961.

KOPAL, Z., Gravitational heating of the moon, *Icarus*, **1,** 412–421, 1963.

LOMNITZ, C., Application of the logarithmic creep law to stress wave attenuation in the solid earth, *J. Geophsy. Res.*, **67,** 365–368, 1962.

LONGMAN, I. M., A Green's function for determining the deformation of the earth under surface mass loads, 2, Computations and numerical results, *J. Geophys. Res.*, **68,** 485–496, 1963.

LOVE, A. E. H., *A Treatise on the Mathematical Theory of Elasticity*, 4th ed., 643 pp., Cambridge University Press, 1927; republ. Dover Publications, New York, 1944.

MACDONALD, G. J. F., Calculations on the thermal history of the earth, *J. Geophys. Res.*, **64,** 1967–2000, 1959.

MACDONALD, J. R., Theory and application of a superposition model of internal friction and creep, *J. Appl. Phys.*, **32,** 2385–2398, 1961.

MUNK, W. H., and MACDONALD, G. J. F., *The Rotation of the Earth*, 323 pp., Cambridge University Press, 1960.

NESS, N. F., HARRISON, J. C., and SLICHTER, L. B., Observations of the free oscillations of the earth, *J. Geophys. Res.*, 621–630, 1961.

TAKEUCHI, H., SAITO, M., and KOBAYASHI, N., Statical deformations and free oscillations of a model earth *J. Geophys. Res.*, **67,** 1141–1154, 1962.

UREY, H. C., ELSASSER, W. M., and ROCHESTER, M. G., Note on the internal structure of the moon, *Astrophys. J.*, **129,** 842–848, 1959.

TABLE I.—*Melting Points of Ultrabasic Rocks*

Material	Initial melting point (°C)	Source
Stoney meteorites (average of 6; mainly chondrites)	1180°[a]	Alexeyeva (1958).
Peridotite (synthetic mixture similar to Nockolds' average peridotite (1954).	1250°	Harris (personal communication).
Serpentine	1300° at P_{H_2O}=500–1000 bars	Clark and Fyfe (1961).

[a]It is interesting to compare the temperature of complete melting of these meteorites (1350°) with the melting point of diopside (1390°C).

system at a temperature lower than the melting points of the individual phases. Ringwood (1962a) has explained this, with reference to the Clausius–Clapeyron equation ($dT/dP=\Delta V/\Delta S$), as due to the fact that the entropy of melting will be increased by the entropy of mixing, while the volume change should be essentially additive. The examples of this behavior cited by Ringwood, namely the values of dT/dP for the MgO–Al_2O_3–SiO_2 systems and for basalt, are about half the values for related individual minerals. It is clear, then, that the use of single mineral fusion curves is an oversimplification which will probably give upper limits for the slope of the fusion curve of rocks.

The initial part of the lunar fusion curve is the melting point at zero pressure of material of chondritic composition. Values for related melting points available in the literature are listed in Table I.

On the basis of these data, we shall assume an initial value of 1250°C for the fusion curve of the lunar material. No specific allowance has been made for the possible lowering of the fusion curve by water pressure; the results reported by Clarke and Fyfe indicate that this would not be very great.

There have unfortunately been no experimental determinations of dT/dP for chondrites nor for ultrabasic rocks. However, fusion curves are available for some of the minerals of such rocks, and are listed in Table II.

In view of the nature of melting in multicomponent systems, discussed previosuly, a value 5°/1000 bars seems a reasonable figure for the slope of the chondritic fusion curve; 10°/100 bars is probably a safe upper limit. These values will be considered constant, although, as Verhoogen points out (1956), they probably decrease somewhat at depth because of the greater compressibility of liquids. (As will be seen, the actual fusion curves will flatten slightly at depth because of the decrease in gravity.)

The effect of phase changes on melting behavior is difficult to evaluate. However, diopside melts congruently out to 50 kb, and Ringwood's (1962) data indicate that the olivine-spinel transition at high temperatures does not occur under about 40 kb. It seems safe to neglect the probable phase

TABLE II.—*Fusion Curves of Potential Lunar Minerals*

Mineral	dT/dP (°C/kb)	Range (kb)	Source
Enstatite	9.5	0–30	Boyd and England (1961).
Diopside	9.2 (avg.)	0–30	Boyd and England (1961).
Forsterite	4.7	"Initial"	Bowen and Schairer (1935).
"Olivine"	3.5	Average in earth's mantle	Verhoogen (1956).
Iron	2.2	0–96	Strong (1959).

TABLE III.—*Pressure-Depth Values*

Depth (km)	50	100	200	300	400	500	600
Pressure (bars)	2680	5270	10,200	14,800	19,200	23,200	26,900

changes in the lunar interior at pressures under 30 kb, corresponding to depths shallower than about 650 km.

Lunar Pressure Gradient

We must now estimate the lunar pressure gradient. Since the strength of the interior of the Moon is probably on the order of 10 to 1000 bars (MacDonald, 1961), which is much smaller than the probable pressures at the depths under discussion, hydrostatic behavior will be assumed. The relation of pressure to depth in a body under hydrostatic conditions is usually found from

$$P = \rho g h,$$

where P = pressure at depth h in material of density ρ, and g = the acceleration of gravity, which is usually considered constant for small values of h. In view of the levels under consideration, however, it seems better to allow for the variation of gravity with depth, which is given by

$$g_x = g_0(r/r_0),$$

where g_x is the acceleration of gravity at a point r/r_0 from the center of the body, and g_0 is the acceleration at the surface. The relation of pressure to depth is then

$$dP/dr = -g_0\rho(r/r_0),$$

which can be integrated to give

$$P = (-g_0\rho/r_0)\int_{r0}^{r} r\,dr$$

$$= (g_0\rho/2r_0)(r_0^2 - r^2).$$

Using this equation, with a value of 3.34 gm/cm^3 for ρ, the pressure-depth values given in Table III can be calculated.

Depth of Magma Generation

Two fusion curves have been calculated, using the above pressures and values of dT/dP of 5°C and 10°C/1000 bars, and superimposed on MacDonald's (1959) diagram showing the present lunar thermal gradients for the "hot" and "cold" Moons (Fig. 1). The "hot" Moon gradient is intersected at about 300 km by both curves, as it is by the diopside curve used by MacDonald. Of more interest, however, is the fact that whereas the diopside curve just misses the "cold" Moon gradient (by a margin which is probably not significant), the fusion curve with the 10°/kb slope intersects it at about 540 km and that with 5°/kb slope at about 440 km.

It is apparent, then, that even in the most conservative case (cold Moon and fusion curve with the highest slope), there should be partial fusion, that is, magma generation, at about 540 km. In the other situations considered, magma will be formed at shallower depths at the present time. In view of the implications of these results, it seems worthwhile to discuss them further: in particular, to inquire as to whether the actual lunar thermal gradient is more closely approximated by the "hot" or the "cold" curves.

There is, of course, one major factor favoring the "cold" Moon, referring in this case to the probable accuracy of this calculated gradient, and

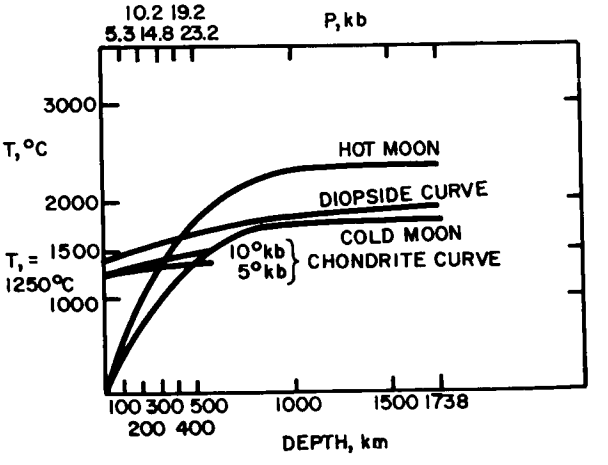


FIGURE 1.—Depth of lunar magma generation. Thermal gradients from MacDonald (1959).

THE RELATION OF TEKTITES TO LUNAR IGNEOUS ACTIVITY

PAUL D. LOWMAN, JR.

Goddard Space Flight Center

The theory that tektites are of lunar origin has frequently been criticized on the grounds that rocks of this chemical composition could not have been formed on the Moon. The purpose of this paper is to show that tektites may have been derived from silicis igneous rocks, specifically rhyolitic tuffs, forming the upper layers of the maria. The essentially igneous nature of tektites is indicated by their bulk composition, their restricted compositional range, and the high probability that they have not been derived from sedimentary rocks by random processes such as meteoritic impact. The differences between tektites and normal igneous rocks probably reflect their formation by extremely high temperature fusion of previously solid igneous rock.

It is shown that if the Moon originally had a chondritic composition, the fusion curve of the lunar material should intersect the thermal gradients computed by G. J. F. MacDonald at the present time at 400 to 500 km depth, depending on the slope of the fusion curve, even if the Moon had originally been at 0°C. Additional heat sources such as the kinetic energy of accretion and capture-induced internal friction would probably have raised the temperature well above the "cold" Moon gradient. Magmas generated by partial fusion of the assumed chondritic material should be basaltic.

It is proposed that the maria are the lunar equivalents of terrestrial lopoliths such as the Bushveld and Sudbury complexes, which are large basins filled mainly with basaltic rock overlain by tuffs, rhyolites, and granophyres. Tektites may have been derived from these silicic rocks by some process involving meteoritic impact.

INTRODUCTION

Despite the great amount of information on the compositions and physical properties of tektites which has been gathered in recent years, the origin of these indisputably real objects is as controversial as ever. A number of facts, such as the absence of cosmogenic isotopes from even the youngest tektites (Viste and Anders, 1962), suggest that if tektites are extraterrestrial, they have probably come from the Moon. A lunar origin is rejected, however, by authorities such as Urey (1958, 1962). One of the chief reasons for this rejection is the belief that there is little if any material on the Moon having a suitable chemical composition for tektite formation, or in more general terms, the belief that the Moon is essentially undifferentiated. [See, for example, discussions by Barnes (1958) and Hawkins (1963).]

The purpose of this paper is to reconcile a lunar origin for tektites with what can reasonably be inferred about lunar geology. The theory will be proposed that igneous processes in the Moon have produced magmas which, after differentiation and extrusion, have formed rocks from which tektites were derived by some process involving meteoritic impact. The actual source areas of the tektite parent materials are suggested to be the maria, which will be shown to be possibly analogous to terrestrial lopoliths.

The general question of terrestrial vs. extraterrestrial origin for tektites will not be discussed here; arguments for each theory are given by the authors cited.

THE IGNEOUS NATURE OF TEKTITES

The theory that tektites are fused sedimentary rocks or soils, as proposed by Barnes (1940), Schwarcz (1962), and others, was necessitated by

the fact that tektites are in many characteristics strikingly different from igneous rocks. Properties peculiar to tektites include the virtually complete absence of crystals, the flow structure, the flanges (of australites), the highly reduced state of the iron, the low water content, and the high $(MgO+FeO)/(Na_2O+K_2O)$ ratio.

Although these characteristics seem to set tektites off from the igneous rocks, numerous others point to an ultimately igneous derivation. Lovering (1960) demonstrated that tektites might be produced from acidic igneous rock such as granophyre by extremely high temperature fusion, which could, by selective volatilization, produce some of the chemical peculiarities referred to above. Lowman (1962) suggested that the functional variation of the various oxides in tektites with silica content resembled lines of magmatic evolution. In a more direct approach, Cuttitta, Carron, Fletcher, and Chao (1962) found the major and minor element compositions of 9 analyzed philippinites and bediasites to resemble most closely rhyolites and granophyres.

Indirect evidence of the igneous nature of tektites is to be found in the high probability that they are not fused sediments. All of the mechanisms proposed thus far for such fusion, such as the impact of comets or meteorites, are essentially random selection processes which must operate on the wide variety of materials exposed on the Earth's surface. However, tektites show remarkable restriction in several chemical properties, such as the isotopic ratios of strontium (Schnetzler, 1961) and oxygen (Taylor and Epstein, 1962), and bulk composition (Schnetzler, 1961; Lowman, 1962). None of the authors cited feels that these properties can be explained as the result of random fusion of terrestrial sediments.

The peculiar physical characteristics of tektites, frequently considered evidence against an igneous origin, have been explained as the result of tektite formation as ablation droplets from satellite meteorites (Baker, 1959; O'Keefe, 1960) or as impactites (Chao and Littler, 1962).

In summary, it seems likely that, although tektites, as they are, have not been formed by solidification of magma, they have been derived from igneous rocks by some process involving extremely high temperature fusion. The question will now be discussed of whether suitable igneous rocks have been formed on the Moon.

LUNAR MAGMA GENERATION

The first step in this discussion must be to investigate the probability that magma generation has occurred in the Moon, and if so, to estimate its depth. This problem is generally considered that of finding where the geothermal gradient is intersected by the fusion curve of the material in question, in this case, the rock composing the interior of the Moon. It will be assumed that the Moon was formed by the accretion of solid material rather than by condensation and solidification. The composition of the primordial Moon will be assumed to be essentially that of chondritic meteorites.

Lunar Thermal Gradients

We shall use the lunar thermal gradients calculated by MacDonald (1959, 1961) as a basis for the present treatment. He derived gradients for a chondritic Moon with initial temperatures (-4.5 aeons) of 0°C and 600°C. MacDonald's values are similar to those computed by Urey (1960) using slightly different assumptions. It is interesting, in view of the formerly popular concept of the Moon as a cold, rigid body, to note that both he and Urey consider it likely that some partial melting would occur at great depths even in an initially cold (0°C) Moon.

Lunar Fusion Curve

The next step in estimating the depth of lunar magma generation is to find a reasonably likely fusion curve for the interior of the Moon. This problem under terrestrial conditions is complicated by the fact that pressures at even a few hundred kilometers are far beyond the range of experimentally determined fusion curves, forcing the use of approximations such as the Simon equation. The relatively low pressure in the Moon, however, permits the application of available experimental data.

Both MacDonald and Urey used the fusion curves of individual minerals, in particular diopside and olivine. However, as MacDonald points out, melting would begin in a multicomponent

not the initial temperature, which would be only slightly affected by the long-lived isotopes discussed. This is the possibility that the assumed values for the concentrations of the major heat-producing isotopes of potassium, uranium, and thorium are too high. This may result from our incomplete knowledge of these concentrations in meteorites, or in a more general sense, from the degree to which the chondrites represent the composition of the Moon. Evidence for a chondritic composition is given by MacDonald (1959) and against it by Urey (1960). A useful disscusion of the validity of the chondritic Earth model is presented by Ringwood (1962); however, his suggestion that the carbonaceous chondrites represent primordial material (and hence possibly the Moon—P.D.L.) seems contradicted by the studies of DuFresne and Anders (1962) indicating that these meteorites are alteration products of normal chondrites. Further discussion of this problem would be beyond the scope of this paper.

The situation in regard to the probability of the "hot" Moon (initial temperature of 600°C) is somewhat clearer in that there are a number of factors which would tend to increase the initial temperature or to add heat at a later time. These will be discussed in order of their possible importance.

The kinetic energy of accretion will add heat to the primitive Moon. Fish, Goles, and Anders (1959) estimate that the temperature rise from this source would be about 2000°K for objects of lunar size, although they suggest that the temperature would not actually get this high because of radiation cooling. It should be pointed out, however, that they were concerned primarily with temperatures high enough to produce the observed characteristics of meteorites, or around 2000–3000°K. It seems unlikely that, if the Moon accreted in less than 10^8 years, it would cool enough by radiation to bring the temperature down to 273°K (the "cold" Moon). The rate of heat loss is proportional to the fourth power of the absolute temperature, and is therefore relatively low at temperatures of a few hundred degrees; the rate of radiation cooling at 873°K (the "hot" Moon) would be less than 3% the rate of 2000°K.

A second heat source of possible importance in determining the initial temperature is adiabatic compression. MacDonald (1959) suggests a temperature increase of several hundred degrees for the Earth due to this cause. The contribution to an accreting Moon cannot be estimated reliably. However, because of the extreme temperature dependence of radiation cooling, it is unlikely that the accretion of the Moon was isothermal.

Another factor tending to increase the temperature of the Moon's interior would be frictional heat generated by body tides. There would be some heat contributed from this source under any hypothesis of the Moon's origin. If the Moon was formed independently near the Earth, the heat would have been generated by the body tides which have slowed the rotation until the lunar day equals the month. The contribution from this mechanism is negligible: approximately 10^{-5} of the heat generated by long-lived isotopes. However, if the Moon was captured by the Earth at about its present distance, the situation might be radically different. Most of the energy loss involved in going from an escape to an orbital trajectory should be dissipated by internal friction in the Moon. Assuming that all the kinetic energy is changed into heat in the Moon, it can be easily calculated that for a velocity change of 1 km/sec, about 8.6×10^{27} cal (or 120 cal/gm) would be produced. This value is of the same order as the heats of fusion of chondritic minerals; for example, ΔH fusion for enstatite is 147 cal/gm (Rossini et al., 1952). The many uncertainties as to the depth of body tides and other factors make more precise calculations unwarranted, but it is clear that if the Moon was captured, the heat so generated might have dominated all other sources, at least temporarily, possibly raising internal temperatures some hundreds of degrees.

Green (1961) has also pointed out that the difference between the Moon's apogee and perigee distances from the Earth should promote magma generation by pressure release due to body tides. He estimates the present monthly tide to be about $1\frac{1}{3}$ meters; the tides in the past should have been higher and more frequent. In view of the previous discussion, the actual heat contribution from this source would probably be small, but the effects of pressure release and fracture formation could be substantial.

Another source of heat may have been the decay

of short-lived isotopes such as Al^{26}, Be^{10}, Ca^{135}, and V^{236}. If the Moon was formed within 10^8 years after the formation of the elements themselves, as seems possible (see discussions by Kopal, 1962 and Anders, 1962), the heat contributed may have been substantial. MacDonald (1959) estimates a temperature rise, for the Earth, of about 3000°C if there was no heat loss. This would not be the situation, of course, but again the temperature dependence of radiation cooling would probably prevent the temperature from going down to 0°C.

Urey and Donn (1956) have suggested that exothermic reactions between various compounds and free radicals during the accretion process of meteorites and the planets may have caused substantial heating if the accretion occurred when the Sun was still relatively cool. They estimate that a lunar-size primary object might have attained a temperature of 500°C throughout.

In summary, it is clear that a number of sources may have contributed substantial heat to the Moon during its early history, suggesting that the "cold" thermal gradient is probably too low since none of the sources listed here were specifically allowed for in MacDonald's calculations. Before leaving the subject of magma generation, however, it is necessary to note another factor. Partial melting of the deep interior of the Moon can take place either by the simple addition of heat until the initial melting point is reached or by the lowering of the melting point by pressure release so that the thermal gradient is intersected, as Green (1961) has pointed out. Yoder (1952) and Uffen (1959) have suggested that such pressure release can occur during tensile or compressive fracture. There is abundant evidence of faulting on the Moon. Magma generation by pressure release might be expected along such faults, regardless of whether they were formed by tension, compression, or shearing, and the faults themselves could provide vents for the magma.

Time of Magma Generation

It is possible, with the use of diagrams presented by MacDonald (1961) showing the variation of temperature with time for various depths in the Moon, to estimate when magma would form. The following times are for a depth of 500 km. For the "hot" Moon, melting should begin at approximately 1.0 aeons (after the origin of the Moon) if the fusion curve had a 10°/kb slope, and at 1.2 aeons for a 5°/kb slope. For the "cold" Moon, melting should begin at 4.0–4.5 aeons for a 10°/kb slope and at 2.7 aeons for a 5°/kb slope.

These ages are essentially a by-product of the present discussion and of course depend on the correctness of the thermal calculations. However, the possibility that magma generation may have begun relatively late in the Moon's history, if it was initially "cold," tends to remove the difficulty Schnetzler (1961) finds in explaining the Sr^{87}/Sr^{86} ratios in tektites. Schnetzler suggests that the North American and Southeast Asia tektites were chemically differentiated from a common source about 175 million years ago, which would require relatively recent differentiation of the Moon if tektites are of lunar origin. This would be consistent with the low age indicated above for magma generation in a "cold" Moon.

Magma Movement

The problem of whether lunar magma could reach the surface requires attention for two reasons. First, it is obvious that lunar igneous rocks proposed as tektite source material must be extrusive, because plutonic rock is unlikely to be exposed over large areas. Second, the depth of potential magma generation indicated by this study (400–500 km) is considerably greater than thought likely in the Earth; for example, Turner and Verhoogen (1960) suggest 100 km as the approximate depth of formation of basaltic magma. (It should be noted, however, that this figure was based on thermal considerations, such as lava temperatures, and hence should not be extrapolated directly to the Moon.)

Wahlstrom (1950) lists several major causes of magma motion, such as gravitative pressure on the magma reservoir, tectonic movements, and gas pressure, and a minor cause, tidal kneading. It is likely that these would all be about as effective in the Moon as in the Earth (assuming them to be operative at all, which may not be the situation for some types of tectonic movement); in fact, tidal kneading might be more effective because the Moon is much smaller than the Earth.

It would seem at first glance that gravitative pressure would be less effective in the Moon (see,

for example, the discussion by Barnes, 1958). However, Turner and Verhoogen (1960) point out that the governing condition for equilibrium in a liquid column is

$$d_1 h = d_2 H,$$

where d_1 and d_2 are densities of the liquid and mantle respectively, h, the depth of origin of the magma, and H the total height of the magma column. This means that the governing factor is *relative* density of the magma; the factor of $\frac{1}{6}$ for the lower lunar gravity will cancel out. Fish, Goles, and Anders (1959) reached a similar conclusion regarding the separation of phases in asteroid-size bodies. (An additional interesting result of their calculations was the conclusion that fractional crystallization and crystal settling would also take place even in weak gravitation fields.) There seems to be no *a priori* reasons, then, why magma formed in the Moon would not reach the surface if suitable pathways are available. A remaining question is whether such pathways actually extend to the probable depth of magma generation.

Deep-focus earthquakes show that failure by fracture occurs, in the Earth, to a depth of about 700 km, or well below the depth where most magma is believed to originate. (Wilson, 1957, suggests that andesitic magmas form at these depths, on the basis of their association with deep-focus earthquakes.) This fact by itself suggests that fractures could extend to a similar level in the Moon, since temperatures probably do not rise as rapidly. However, there is objective evidence indicating the possible existence of deep fractures. This is found in the numerous faults visible on the lunar surface, such as the Straight Wall (over 100 km long) and the Altai Scarp (over 300 km long), as well as the fracture system radiating from Mare Imbrium for a distance of over 1100 km. The depth to which these fractures extend is of course unknown, but it seems reasonable to expect them to reach the depth of magma generation. For example, the well-known dark areas at bends in the Alphonsus rille system suggest that some form of vulcanism has occurred along these fractures, implying that they have tapped a magmatic zone.

Nature of Lunar Primary Magma

Considering the uncertainty about the origin of terrestrial magmas whose products have been exhaustively studied at first hand, it might seem that any inquiry as to the nature of lunar magmas would be baseless speculation. However, since this study is intended to answer a primarily chemical question on the basis of an assumed chemical composition; such an attempt is both necessary and possible.

A major factor governing the composition of the magmas formed by partial melting is the composition of the parent rock, assumed to be chondritic in the case of the Moon. There is no experimental data on the fusion products of chondrites, so that the following discussion must rest on the similarity of chondrites to peridotites, which is illustrated by the following comparison (Table IV).

In chondrites the silicate phase is about 81% (by weight) of the total, with metal (iron and nickel) and sulfides comprising most of the remainder. A comparison of melting points at 1 atm suggests that the sulfides would melt first, silicates next, and iron last. The textures of meteorites show that these three phases would be mutually insoluble, and judging from the calculations of Fish, Goles, and Anders (1959), the iron and sulfides would sink toward the center. It seems safe

TABLE IV.—*Chemical Compositions of Chondrites and Peridotites*

	Wt. % of						
	SiO_2	MgO	FeO	Al_2O_3	CaO	Na_2O	K_2O
Peridotites (Nockolds, 1954)	43.0	34.3	12.4	4.0	3.5	0.6	0.2
Silicate phase of chondrites (Urey and Craig, 1953)	47.0	29.5	15.4	3.1	2.4	1.2	0.2

to assume that at depths of a few hundred kilometers in the Moon the nonsilicate phases would have relatively little effect on the melting of the silicates.

There is considerable field and experimental evidence indicating that the first-formed liquid from fusion of peridotite is basaltic in composition (Harris and Rowell, 1960 and personal communication; Turner and Verhoogen, 1960), and that oversaturated basalts have been generated at depths of around 100 km in the Earth's mantle. The conditions peculiar to the Moon may cause magmas to evolve along different lines than their terrestrial counterparts, as pointed out by Walter (1962). However, to the extent that the upper mantle of the Earth, the interior of the Moon, and chondrites are similar, lunar primary magmas will be basaltic. A discussion of this similarity would be beyond the scope of this paper.

Before leaving the subject of lunar magma generation, an interesting possibility should be pointed out. The experimental work cited previously, in particular that of Harris and Rowell (personal communication) and Alexeyeva (1958) shows that the temperature interval between the initial and complete melting of ultrabasic rocks is relatively short, apparently between 100° and 200°. As shown in Fig. 1, the thermal gradient for the "cold" Moon exceeds the 5°/kb fusion curve by 200° at about 650 km. This condition, together with the probable suppression by pressure of incongruent melting in minerals such as enstatite (Boyd and England, 1961), might permit the generation of ultrabasic magmas. However, such magmas are rarely, if ever, formed in the Earth, and it seems unlikely that conditions in the Moon would be different enough to permit their formation there. Moreover, several independent lines of evidence, such as the luminescence of the Aristarchus rays (Kozyrev, 1956), the polarization of moonlight (Wright, 1927), and the dielectric constant of the surface as calculated from radar returns (Senios and Siegel, 1960) indicate that the surface material is mainly siliceous powder, rather than ultrabasic rock.

TEKTITE SOURCE AREAS

It has been shown that tektites have igneous affinities, and that there have probably been substantial quantities of basaltic magma generated in the Moon and erupted on the surface. However, tektites can hardly have been formed from basalt by any of the fusion mechanisms so far suggested; it is apparent that a clear connection between tektites and the Moon has not yet been demonstrated.

The hypothesis is therefore proposed (see also Lowman, 1962) that the lunar maria are lopoliths, and that tektites have been derived by some process involving meteoritic impact from the silicic igneous rocks overlying these lopoliths. In support of this hypothesis, a comparison of terrestrial lopoliths and the maria will be presented.

The term "lopolith" was originally coined by Grout (1918) for large lenticular intrusions with depressed floors, and consisting mainly of basaltic rocks with subordinate volumes of ultramafic and granitic rocks. The best examples are the Duluth, Bushveld, and Stillwater lopoliths. The Sudbury basin may also be a lopolith, although the suggestion has resently been made (Thomson, 1957) that it is actually a ring-dike. An important characteristic of the large lopoliths is their location in essentially nonorogenic regions; although they have all been deformed after their formation, they are not associated with folded mountain belts, as are batholiths. Daly (1928) suggested that the Bushveld lopolith was largely extrusive; Hamilton (1960) presented evidence that the Sudbury and Wichita lopoliths are also extrusive, or at least are roofed with silicic extrusives such as tuffs and rhyolites.

The lunar maria have been generally considered to be plateau basalts. However, Shoemaker and Hackman (1961) suggest that either basalt sheets or ignimbrite could produce the flatness and smoothness of the maria. O'Keefe and Cameron (1962) propose that the maria are subsidence structures which are probably overlain by welded tuff (ignimbrite) and filled with sufficient denser rock to induce isostatic subsidence. Although O'Keefe and Cameron did not specifically suggest that the maria were lopoliths, it is apparent that their description in general fits the lopoliths fairly well. This similarity is emphasized by the comparison of major characteristics of these structures presented in Table V.

TABLE V.—*Comparison of Lopoliths and Maria*

	Lopoliths*	Maria
Size (km)	Bushveld complex, 200×500 Sudbury basin, 60×25 Duluth lopolith, 230 km exposed along strike.	M. Imbrium, 1100. M. Crisium, 500×400. (O. Procellarum omitted because of probable composite nature.) Grimaldi, 130 (Large crater filled with mare-type material).
Shape	Elliptical; essentially closed synclines. Thickness ranges from several km to 15(?) for the Bushveld complex.	Circular or elliptical; outcrop of mare material frequently irregular. Structure and topography syclinal. Maximum thickness unknown, but average thickness of Procellarian System estimated at 1.1 km by Marshall (1961).
Rock types	Basalt, diabase, and gabbro dominant. Subordinate quantities of ultramafic and granitic rocks, including rhyolites, granophyres, and tuffs. Silicid rock layer a few thousand meters thick in places, although generally removed by erosion or covered.	Exposed portions light-colored, porous material, possibly welded tuffs; underlain by rocks heavy enough to induce isostatic subsidence.

*Descriptions of lopoliths from papers by Daly (1928), Hess (1960), and Hamilton (1956, 1960).

A further comparison between the maria and lopoliths is illustrated in Fig. 2, in which the approximate outline of the Bushveld Complex is superimposed on Mare Crisium. It is apparent from this diagram and the tabulated characteristics that although there is wide overlap in size, the large maria, such as M. Imbrium and M. Serenitatis, are considerably bigger than the lopoliths. This can be explained, in reference to the O'Keefe–Cameron theory as follows. The terrestrial lopoliths are all Precambrian, and have been extensively eroded; their former silicic caps in particular have been largely removed. On the Moon, however, the absence of erosion in the usual sense has probably preserved the original shape and extent of the maria. In addition, as O'Keefe and Cameron point out, the ash flows which deposited the welded tuffs supposed to comprise the upper parts of the maria would have been much more mobile on the Moon than they would have been on the Earth, because the lower gravity would greatly increase the effectiveness of fluidization. The result of these two effects has probably been the extension of the silicic differentiates of the maria (the welded tuffs) well beyond the limits of the mafic central portions, which would correspond to the existing terrestrial lopoliths. The size of the central mare basins is probably fairly close to that of the lopoliths; there is at any rate no such scaling problem as that encountered in explaning the large lunar craters as calderas (Green, 1961; Shoemaker, 1962).

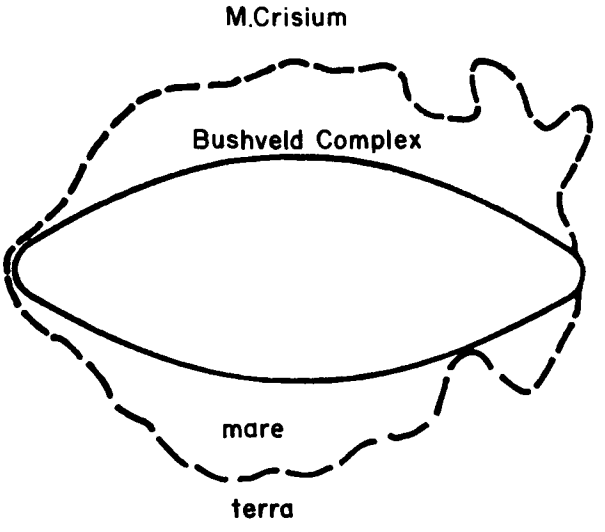


FIGURE 2.—Comparison of M. Crisium and Bushveld Complex. Bushveld outline from Daly (1928).

Another characteristic which may link tektites to the silicic differentiates of the maria is the $(FeO+MgO)/(K_2O+Na_2O)$ ratio. As shown in

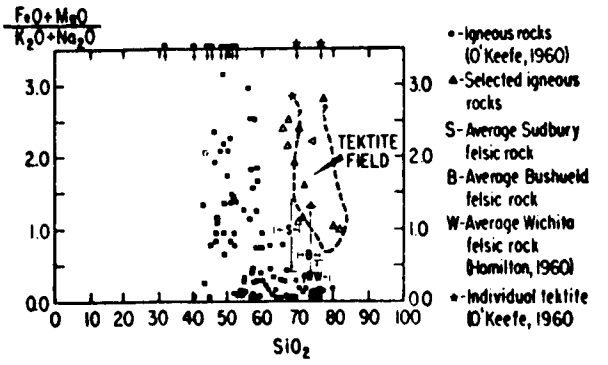


FIGURE 3.—$(FeO+MgO)/(K_2O+Na_2O)$ ratios for tektites and igneous rocks.

Fig. 3, the average ratios for the lopolith rocks bridge the gap between the common igneous rocks and tektites. If there were substantial loss of alkalis from lopolith differentiates by volatilization, they would move into the tektite field. A similar conclusion was reached by Lovering (1960) on the basis of experimental studies, although it is not known if he worked with lopolith rocks.

It should be stressed, in basing any comparison on the supposed characteristics of terrestrial lopoliths, that the nature and evolution of these structures is far from being fully understood. In addition to the doubts as to the lopoliths structure of the Sudbury Basin raised by Thomson (1957), there is considerable controversy about the origin of the granitic igneous rocks overlying the lopoliths. Hamilton (1960) suggest that they were formed by differentiation of the basaltic magmas which formed the main bodies of the lopoliths. Hess (1960), on the other hand, argues that the relative abundance of the granitic rocks is better explained as the result of partial fusion of the continental crust (above the M discontinuity) by the rising isogeotherms responsible for the entire series of igneous events involved in lopolith formation. An additional note of controversy has recently been injected into the lopolith literature by Dietz (1962), who suggests that both the terrestrial lopoliths and the maria were triggered by the impact of extremely large meteorites.

Which, if any, of these viewpoints is more nearly correct cannot be decided by the writer. In view of the relative rates of progress of geology and astronautics, however, it is possible that the origin of terrestrial lopoliths will be elucidated by field studies of the maria before the converse can be accomplished.

CONCLUSIONS

It has been shown that tektites are probably altered igneous rocks, and that suitable igneous rocks may cover large areas of the visible face of the Moon. The theory that tektites have been derived from the maria by some process involving meteoritic impact is supported by the existence of several large craters, such as Copernicus, which are convincingly demonstrated (Shoemaker, 1962) to have been formed by the impact of meteorites on the maria.

The picture of lunar evolution presented here has some obviously controversial features. For example, if the more conservative "cold" Moon is assumed, it follows that the maria (including the smaller bodies such as Grimaldi and Mare Undarum) represent practically all of the differentiation which has occurred in the Moon. This in turn implies that the lighter-colored terrae are largely chondritic and the maria silicic, at least superficially, or precisely the reverse of the generally accepted concept of silicic terrae and basaltic maria. However, O'Keefe and Cameron (1962) have presented evidence that the Moon is largely in isostatic equilibrium and that the terrae are therefore probably granitic. If their conclusion is correct, there has been more differentiation than that involved in maria formation.

On the other hand, the mechanism of differentiation proposed here, namely partial melting by largely radiogenic heat along deep fractures, is analogous to processes which probably take place in the Earth, and requires no radical assumptions, such as the impact of extremely large meteorites relatively late in the Moon's development. In addition, to the extent that the melting was partial and local, this theory tends to answer the problem raised by MacDonald (1962) of "how differentiation can proceed without the Moon's passing through a molten stage."

An advantage of the theory that the maria are lopoliths is that, if Hamilton's concepts of lopolithis structure and development are correct, a mechanism is available by which the Moon may have been differentiated without the intervention of sedimentary processes. The processes by

which the Earth's granitic crust was formed are probably dependent in considerable degree on the accumulation of great quantities of sediment in geosynclines, even though most granites may be magmatic (see discussions by Tuttle and Bowen, 1958: and Walton, 1960). However, terrestrial lopoliths are clearly not part of the usual calc-alkaline igneous association of orogenic regions, and are therefore not tied firmly to sedimentation.

A question unresolved by this discussion is whether the maria are in any way related to impact, as suggested by several authors (Baldwin, 1949; Urey, 1952). Dietz (1962) also holds this opinion, and further suggests that the terrestrial lopoliths, such as the Sudbury basin, are the result of the impact of extremely large meteorites. A detailed discussion of Dietz's theory would be beyond the scope of this paper. However, it should be pointed out that the petrographic association typical of lopoliths is also found in the Great Dike of Rhodesia, whose shape (about 300 by 5 miles) virtually excludes an impact origin. It seems safe to say that if impact was involved in lopolith evolution, it role was limmited to initiating a complex series of tectonic and igneous events, perhaps by fracturing the target rock to great depths.

In summary, it is suggested that the concept of the maria as lunar lopoliths correlates a number of previously unconnected data, provides a basis for a new engineering model of possible lunar surface conditions, and is specific enough to be tested by observational and field studies.

ACKNOWLEDGMENTS

A number of people gave the writer valuable assistance in this study. Dr. Peter Harris made available the unpublished preliminary results of his experiments on peridotite fusion. Drs. W. Hamilton, H. Hess, H. Yoder, and F. Boyd read an early version of the manuscript and suggested many possible improvements. The writer had valuable discussions of the subject matter with Drs. J. O'Keefe, L. Walter, E. Shoemaker, J. Green, R. Smith, V. Barnes, and Mrs. W. Cameron.

Dr. V. Osvalds, Acting Director of Leander McCormick Observatory, kindly permitted the writer to use the Observatory's telescope in company with R. J. Hackman of the U.S. Geological Survey, whose suggestions on lunar structure were very helpful.

Thanks are also due to Dr. Z. Kopal for a translation of Kozyrev's paper on luminescence of the lunar surface.

I am very grateful to these people for their generosity, and would like to stress that the opinions expressed are not neccessarily endorsed by any of those mentioned.

REFERENCES

Alexeyeva, K. N., Physical properties of stoney meteorites and their interpretation in the light of hypotheses of meteoritic origin, *Meteoritics*, **16,** 66–77 (1958).

Allen, C. W., "Astrophysical Quantities," Oxford Univ. Press, London and New York (1955).

Anders, E., Meteorite ages, *Rev. Mod. Physics*, **34,** 287–325 (1962).

Baker, G., Tektites, *Victoria Natl. Museum, Mem.*, **23** (1959).

Baldwin, R., "The Face of the Moon," Univ. of Chicago Press, Chicago, Illinois (1949).

Barnes, V. E., North American tektites, *Univ. Texas Publ.*, **3945** (1940).

———, Properties of tektites pertinent to their origin, *Geochim. et Cosmochim. Acta*, **14,** 267–278 (1958).

Bowen, N. L., "The Evolution of the Igneous Rock," reprint ed. Dover Press, New York, 1955 (1928).

———, and Schairer, J. F., The system MgO–FeO–SiO_2, *Am. J. Sci.*, **29,** 151–217 (1935).

Boyd, F. R., and England, J. L., Melting of silicates at high pressures, Ann. Rept. Director Geophys. Lab., *Carnegie Inst. Wash. Year Book*, **60,** 113–125 (1961).

Chao, E. C. T., and Littler, J., Additional evidence for the impact origin of the Ries Basin, Bavaria, Germany (abstract), *Program Ann. Meeting Geol. Soc. Am., Houston, Texas* (1962).

Chayes, F., Numerical petrography, Ann. Rept. Director Geophys. Lab., *Carnegie Inst. Wash. Year Book*, **60,** 158–165 (1961).

Clark, R. H., and Fyfe, W. S., Ultrabasic liquids, *Nature*, **191,** 158–159 (1961).

Cuttitta, F., Carron, M. K., Fletcher, J., and Chao, E. C. T., "Chemical Composition of Bediasites and Philippinites," Astrogeologic Studies, Semiann. Progress Rept., Feb. 26, 1961, to Aug. 24, 1961, U.S. Geological Survey (1962).

Daly, R. A., Bushveld igneous complex of the Transvaal, *Bull. Geol. Soc. Am.*, **39,** 703–768 (1928).

Dietz, R. S., Meteoritic impact origin of the Moon's surface features, *J. Geol.*, **54,** 359–374 (1946).

———, Sudbury structure as an astrobleme (abstract), *Trans. Am. Geophys. Union*, **43**(4), 445–446 (1962).

DUFRESNE, E. R., and ANDERS, E., On the chemical evolution of the carbonaceous chondrites, *Geochim. et Cosmochim. Acta*, **26,** 1085–1114 (1962).

FISH, R. A., GOLES, G. G., and ANDERS, E., "The Record in the Meteorites, III." Rept. EFINS–59–43, Enirco Fermi Inst. for Nuclear Studies, University of Chicago (1959).

FRIEDMAN, I., THORPE, A., and SENFTLE, F., Comparison of the chemical composition and magnetic properties of tektites and glasses formed by fusion of terrestrial rocks, *Nature*, **187,** 1089–1092 (1960).

GREEN, J., "The Geology of the Lunar Base." Rept SID 61–358, Space Sci. Lab., North Am. Aviation, Inc., Downey, California (1961).

GROUT, F. F., The lopolith, an igneous form exemplified by the Duluth gabbro, *Am. J. Sci.*, **46,** 516–522 (1918).

HAMILTON, W., Precambrian rocks of Wichita and Arbuckle Mountains, Oklahoma, *Bull. Geol. Soc. Am.* **67,** 1319–1330 (1956).

———, Silicic differentiates of lopoliths, *Rept. 21st Intern. Geol. Congr., Copnhagen*, Part 13, 59–67 (1960).

HAWKINS, G. S., A study of tektites, *J. Geophys. Res.*, **68,** 895–910 (1963).

HARRIS, P., and ROWELL, J. A., Some geochemical aspects of the Mohorovicic discontinuity, *J. Geophys. Res.*, **65,** 2443–2459 (1960).

HESS, J. J., Stillwater igneous complex, *Geol. Soc. Am. Mem.*, **80** (1960).

JEFFREYS, SIR H., "The Earth," 3rd ed., Cambridge Univ. Press, London and New York (1952).

KOPAL, Z., "Internal Structure of the Moon." Lecture delivered at the Lunar Missions Meeting, Cleveland Am. Rocket Soc. Preprint 2455–62 (1962).

KOZYREV, N. A., Luminescence of the lunar surface and intensity of corpuscular radiation of the Sun, *Izv. Crimean Astrophys. Obs.*, **16,** 148 (1956).

LOVERING, J. F. (1960). High-temperature fusion of possible parent materials for tektites, *Nature*, **186,** 1028–1030 (1960).

LOWMAN, P. D., Tektites vs. terrestrial rocks: a comparison of variance in compositions, *Geochim. et Cosmochim. Acta*, **26,** 561–579 (1962).

———, The relation of tektites to lunar igneous activity (abstract), *J. Geophys. Res.*, **67,** 1646 (1962).

MACDONALD, G. J. F., Calculations on the thermal history of the Earth, *J. Geophys. Res.*, **64,** 1967–2000 (1959*a*).

———, Chondrites and the chemical composition of the Earth, *in* "Researches in Geochemistry" (P. H. Abelson, ed.), p. 476–494, Wiley, New York (1959*b*).

———, Stress history of the Moon, *Planet Space Sci.*, **2,** 249–255 (1960).

———, Interior of the Moon, *Science*, **133,** 1045–1050 (1961).

———, On the Internal Constitution of the Inner Planets, *J. Geophys. Res.*, **67,** 2945–2974 (1962).

MARSHALL, C. H., Thickness of the Procellarian system, Letronne region of the moon (abstract) *Program Ann. Meeting Geol. Soc. Am., Cincinnatio, Phio*, p. 101A (1961).

MUNK, W. H., and MACDONALD, G. J. F., "The Rotation of the Earth," Cambridge Univ. Press, London and New York (1960).

NOCKOLDS, S. R., Average chemical composition of some igneous rocks, *Bull. Geol. Soc. Am.*, **65,** 1007–1032 (1954).

O'KEEFE, J. A., The origin of tektites, *NASA Tech. Note* D–490 (1960).

———, and CAMERON, W. S., Evidence from the Moon's surface for the production of lunar granites, *Icarus*, **1,** 271–285 (1962).

RINGWOOD, A. E., A model for the upper mantle, *J. Geophys.*, **67,** 857–867 (1962).

ROSSINI, F. D., WAGMAN, D. D., EVANS, W. H., LEVINE, S., and JAFFE, I., Selected Values of Chemical Thermodynamic Properties, Part I, *Natl. Bur. Standards (U.S.) Circ.* **500** (1952).

SCHNETZLER, C. C., The composition and origin of tektites, Ph.D. Thesis, Massachusetts Institute of Technology.

SCHWARCZ, H. P., A possible origin of tektites by soil fusion at impact sites, *Nature*, **194,** 8–10 (1962).

SENIOR, T. B. A., and SIEGEL, K. M., A theory of radar scattering by the Moon, *J. Res. Natl. Bur. Standards*, **64D,** 3 (1960).

SHOEMAKER, E. M., Penetration mechanics of high velocity meteorites, illustrated by Meteor Crater, Arizona, *Rept. 21st Intern. Geol. Congr., Copenhagen*, Part 18, 418–434 (1960).

———, Interpretation of lunar craters, *in* "Physics and Astronomy of the Moon" (Z. Kopal, ed.), p. 283–359, Academic Press, New York (1962).

———, and HACKMAN, R. J., "Stratigraphic Basis for a Lunar Time Scale," Astrogeologic Studies, Semiann. Progress Rept., Aug. 25, 1960, to Feb. 25, 1961, U.S. Geological Survey (1961).

SPURR, J. E., "Geology Applied to Selenology," Science Press, Lancaster, Pennsylvania (1944).

STRONG, H. M., The experimental fusion curve of iron to 96,000 atmospheres, *J. Geophys. Res.*, **64,** 653–659 (1959).

TAYLOR, H., and EPSTEIN, S., Oxygen isotope studies on the origin of tektites, *J. Geophys. Res.*, **67,** 4485–4490 (1962).

THOMSON, J. E., Geology of the Sudbury basin, *Ann. Rept. Ontario Dept. Mines*, **65,** 1–56 (1957).

TURNER, F., and VERHOOGEN, J., "Igneous and Metamorphic Petrology," McGraw-Hill, New York (1960).

TUTTLE, O. F., and BOWEN, N. L., Origin of granite in the light of experimental studies in the system $NaAlSi_3O_8$–$KAlSi_3O_8$–SiO_2–H_2O, *Geol. Soc. Am. Mem.*, **74** (1958).

UFFEN, R. J., On the origin of rock magma, *J. Geophys. Res.*, **64,** 117–122 (1959).

UREY, H. C., "The Planets," Yale Univ. Press, New Haven, Connecticut (1952).

———, Origin of Tektites, *Nature*, **181,** 1457 (1958).

———, Lines of evidence in regard to the surface of the Moon, *in* "Space Research—Proceedings of the First International Space Science Symposium" (H. Kallmann-Bijl, ed.), p. 1114–1122, North-Holland, Amsterdam (1960).

———, and CRAIG, H., The composition of the stone meteorites and the origin of the meteorites, *Geochim. Cosmochim. Acta*, **4**, 36–82 (1953).

———, and DONN, B., Chemical heating for meteorites, *Astrophys. J.*, **124**, 307–310 (1956).

VERHOOGEN, J., Temperatures within the Earth, *in* "Physics and Chemistry of the Earth" (L. H. Ahrens, K. Rankama, and S. K. Runcorn, eds.), Vol. I, Pergamon Press, New York (1956).

VISTE, E., and ANDERS, E., Cosmic-ray exposure history of tektites (abstract), *J. Geophys. Res.*, **67**, 1661 (1962).

WAHLSTROM, E. E., "Theoretical Igneous Petrology," Wiley, New York (1950).

WALTER, L. S., Lunar magmatic differentiation (abstract), *Trans. Am. Geophys. Union*, **4**, 464 (1962).

WALTON, M., Granite problems, *Science*, **131**, 635–645 (1960).

WILSON, J. T., The crust, *in* "The Earth and Its Atmosphere" (D. R. Bates, ed.), p. 48–73, Basic Books, New York (1957).

WRIGHT, F. E., Polarization of light reflected from rough surfaces, with special reference to light reflected by the Moon, *Proc. Natl. Acad. Sci. U.S.*, **13**, 535–540 (1927).

YODER, H. S., Change of melting point of diopside with pressure, *J. Geol.*, **60**, 364–374 (1952).

THE STRUCTURE AND STRENGTH OF THE INNER PLANETS

GORDON J. F. MacDONALD

*Goddard Space Flight Center**

Recent developments in seismology and geodesy provide data for an improved model of the earth's interior. This report examines the internal structures of the moon, Mars, Venus, and Mercury in the light of what is known about the internal constitution of the earth.

A review of the seismic determination of the elastic constitution of the earth's mantle, based on new results on the stability of silicates at high pressures, leads to the following interpretation that the rapid increase of elastic wave velocity beginning at a depth of 200 km depends on the olivine-spinel transition and the breakdown of silicates to oxides. Preliminary calculations of the stability field of periclase (MgO) and stishovite (very dense SiO_2) relative to olivine ($MgSiO_4$) indicate that the oxides are stable at pressures greater than 100,000 to 150,000 bars. The oxide transition produces a change in volume of about 20 percent.

The earth's gravitational figure, as obtained from satellite orbits, is used to estimate the possible deviations from hydrostatic equilibrium in other planets. The earth's response to tidal excitation provides information regarding the deviations of elasticity within the earth. The near coincidence of the present rate of heat production of a chondritic earth and the present surface heat flow is discussed as a limiting condition on the earth's internal thermal structure.

Observations of the moon's orbital and rotational motions provide data on its gravitational figure. The calculation of its thermal structure shows that a model moon with uniform radioactivity and chondritic composition is inconsistent with the present figure. The inferred strength of the moon requires either that the radioactivity be substantially less than that of chondrites or that heat sources are concentrated in the outer layers. The problem of differentiation without melting is noted. The average lunar material has a radioactivity perhaps one-half or less that of chondritic materials.

The astronomical information on the mean density and gravitational figure of Mars is critically examined. If the mean radius of Mars is taken to be 3310 km, the planet must have a surface density of 3.8–3.9 gm/cm^3 and be nearly homogeneous. If Mars were homogeneous and had a radioactivity equal to that of chondrites, the interior would be molten and large scale gravitational differentiation would be expected. Since this differentiation is not apparent in the gravitational data, it is concluded that the radioactive composition of Mars differs from that of chondrites.

The internal structures of Mercury and Venus are examined in terms of their inferred rotational history.

It is finally concluded that the inner planets differ in their abundances both of the heavy elements and of potassium, uranium, and thorium. Chondritic meteorites may provide a satisfactory chemical model for the earth, but not for the other inner planets, with the possible exception of Venus. This qualification is due to our ignorance of the internal constitution of Venus.

Section 1

INTRODUCTION

A number of developments in the past few years have stimulated interest in the nature and origin of the planets. In the nineteenth and the early twentieth centuries, investigation of the planetary system utilized the concepts and methods of celestial mechanics. The past 20 years have seen several major breaks with this tradition. Urey emphasized the chemical aspects of the problem of planetary origins and, using thermochemical considerations, provided powerful re-

*This work was done while the author was on leave from the Institute of Geophysics and Planetary Physics, University of California.

straints on speculation.[1] Alfvén insisted that the magnetic fields play a role comparable in importance with that of the gravitational fields.[2] Alfvén's considerations have been extended and deepened, particularly by Lüst and Schlüter[3] and by Hoyle.[4] Cameron[5] and Fowler, Greenstein, and Hoyle,[6] after the discovery of xenon-129 in meteorites by Reynolds,[7–9] demonstrated the constraints placed by nuclear physics on the early history of the solar system. But by far the greatest impetus to further study of the planets has been the promise of direct exploration of the moon, Mars, and Venus. The advent of planetary probes also provides numerous new tools for this study.

Knowledge of the earth and its interior has grown in the past few years. The determination of the constitution of the earth provides a framework within which the observations of the moon and terrestrial planets can be interpreted. Jeffreys first examined the constitution of the earth-like planets in the light of geophysical knowledge.[10] During the past ten years the understanding of the internal constitution of the earth has increased. The study of the earth's gravitational field with artificial satellites and the investigation of its interior by means of observations of low frequency seismic waves have added a new dimension to this understanding. Although major problems still abound in the theory of the earth's interior, its internal distribution of mass and elasticity are relatively well determined. The question of its thermal regime remains a major problem. The origin of its surface features is closely tied to its anelastic behavior, and little progress has been made in this field. The origin of the geomagnetic field remains a mystery, though magnetohydrodynamic theories developed by Bullard and by Elsasser indicate a promising approach to the answer. This report will review those areas of the study of the earth that have contributed most heavily to the understanding of its internal constitution. The methods, concepts, and techniques applicable to the earth should also apply to the study of the moon and the other inner planets. Therefore, after considering the state of knowledge about the earth, we will discuss the current state of knowledge of the internal constitution of the moon, Mars, Venus, and Mercury. Particular attention will be paid to those problems about which a great deal will be discovered in the early stages of planetary exploration.

The determination of the earth's figure is the principal concern of geodesy. Since the external figure of the earth is the result of the internal mass distribution, a detailed knowledge of the earth's gravitational field sets limits on the possible internal mass distributions. The departure of the earth's gravitational field from that of a homogeneous sphere results in observable perturbations in the motions of the moon and of artificial satellites. The study of the orbits of artificial satellites has been particularly valuable in providing additional information regarding the earth's external potential. The deviations of the earth's figure from that which would be assumed by a fluid body rotating at the same angular velocity provide some measure either of the earth's strength in resisting long-period small-amplitude stress differences or of this strength and the internal dynamic processes that support the figure inequalities.

Gravitational forces of the moon and sun raise tides in the solid body of the earth; and these tidal forces are precisely known. Accurate observation of the earth's response to these forces yields information on the distribution of both elasticity and density within the earth and on anelasticity of the earth at the frequencies of the tidal forces. The variation of the earth's rotation—both in terms of the variation of position of the instantaneous axis of rotation and in terms of the length of day—also yields information regarding the internal elasticity and anelasticity.

The most powerful tool in the investigation of the earth's interior has been a determination of the time needed for elastic waves to travel through the earth. The source of such elastic waves is, for the most part, earthquakes, though in recent years artificial sources have become increasingly important. The determination of the travel time of these waves permits an estimation of their velocity variations, and this can be combined with geodetic data to obtain the distribution of elasticity and density within the earth. In recent years an additional powerful investigational tool has been the observation of the earth's free oscillations. Because free oscillations involve the planet

as a whole, this method reduces the difficulties associated with analyzing many events on different seismographs for several earthquakes.

The variation of elastic wave velocity with depth, when combined with laboratory investigations, permits a first guess as to the temperature distribution in the interior of the earth. A more direct method involves measurements of the heat flowing from the interior through the surface. Locally, this surface heat flux depends on the vertical temperature gradient and thermal conductivity. The *total* surface heat flux is fixed by the internal distribution of radioactive heat sources and of thermal conductivity, and on thermal conditions during the early stages of the earth's formation. Models of the earth involving the distribution of radioactivity and initial thermal conditions can be constructed, leading to the observed total heat flux. These models also give some indication of possible internal temperature distributions. However, the actual distribution remains most uncertain.

A very important aspect of the investigation of the earth's interior concerns the behavior of materials at high pressures and temperatures. In the last few years great progress has been made in simulating deep-earth conditions in the laboratory. Materials that are stable at the earth's surface undergo a vast variety of changes when subjected to the conditions that prevail in the upper 100 km of the earth. It is likely that further changes in the structure of matter will occur in the laboratory experiments as they more closely approach conditions existing in the *deep* interior, where pressures are on the order of 10^6 bars at temperatures on the order of a few thousand degrees. Laboratory data can be combined with seismically derived velocity-depth variation data to construct chemical-physical models of the earth's interior. But the laboratory data do not yet cover a sufficient range, and most models depend heavily on untested theories of the behavior of matter at high temperatures and pressures.

After reviewing the state of knowledge of the earth in the fields considered above, we will examine the present state of ignorance regarding the moon and the other planets. Though it is by far the best known of the nearby bodies, the moon's internal character remains a mystery. The two small Martian satellites provide data on the gravitational figure of the Red Planet. These data are consistent with a number of possible models for the internal constitution of Mars, but a combination of arguments permits some further limits to be placed on the planet's internal character. The natures of Mercury and Venus are even more uncertain. Almost nothing is known about Venus because of its continuous cloud cover. Mercury remains anomalous with its much higher density than the other members of the solar system.

Section 2

THE EARTH'S GRAVITATIONAL FIELD

The earth's geometrical figure and its gravitational field are closely related, since the figure is defined in terms of the mean sea level and the mean sea level is fixed by the ocean surface, which is everywhere perpendicular to the direction of gravity. The dimensions of the earth can be estimated by geodetic surveys. Latitudes and longitudes are found by astronomical observation, and surveys along a meridian determine a length. A comparison of the length with the differences in latitude and longitude fixes the local figure. In recent years large arcs have been surveyed on land and geodetic measurements across oceans have been attempted.

A second way of determining the earth's figure involves the measurement of the gravitational field at various places on the surface. Prior to the advent of satellites this was the principal method of determination. It suffers from the fact that the earth's gravitational field contains power in all harmonics. Field irregularities, which can be described as high-wave-number gravitational noise, make it difficult to estimate the field from scattered surface observations. In recent years statistical methods have been applied to the reduction of the gravitational data. These lead to a far superior description of the field, yet they are still inadequate for the low-wave-number components in comparison with the determination made by using satellites.

On account of the flattening of the earth, the attraction of the earth and moon produces a

couple which induces motion of the earth's rotation axis about the pole of the ecliptic. This motion is known as precession and the precessional constant H is given by

$$H=(C-A)/C,$$

where C is the principal moment of inertia about the axis of rotation and A is the moment of inertia about an axis in the equatorial plane. H can be determined from the period of precession. The observed precession of the equinoxes yields a value of 1/305.3; the principal uncertainty in this value is due to the uncertainty in the mass of the moon.

Artificial satellites offer a powerful method for determining the earth's gravitational field. The orbit of a satellite about an isolated homogeneous spherical earth devoid of an atmosphere would be an exact ellipse. The size and shape of the ellipse would remain constant; the plane of the ellipse would be fixed relative to the stars and the orientation of the ellipse in its own plane would not change. But the earth does have an atmosphere, the sun and moon perturb the motion of a satellite, and, most importantly, the deviations of earth from a homogeneous sphere perturb the ellipse.

The two most important perturbations of the satellite's orbit which are due to the earth's internal density field are the steady rotation of the orbital plane about the earth's axis and the rotation of the major axis of the orbit in its own plane. These rotations are determined by the even order harmonics of the earth's gravitational field. Careful observations of many satellites have fixed the lower order even harmonics. In addition to the secular variation of the satellite orbit, there are long period oscillations due to the odd order harmonics.

Satellite orbits are also perturbed by the atmosphere and the tidal actions of the sun and moon. Fortunately, these effects can be removed for the most part. The main effect of the atmosphere is to make the orbit contract and become more circular; its effect on the rotation of the orbital plane is negligible.

Description of the Exterior Potential

The earth's potential at a distance r from the center of mass of the earth can be expressed as a sum of spherical harmonics:

$$U=\frac{GM}{r}\left\{1-\sum_{n=2}^{\infty}\left(\frac{a_e}{r}\right)^n\left[J_n^0 P_n^0 \sin\theta+\sum_{m=1}^{n}\left(J_n^m\cos m\lambda+K_n^m\sin m\lambda\right)P_n^m\sin\theta\right]\right\}, \tag{1}$$

where G is the gravitational constant, M the mass, and a_e the equatorial radius of the earth ($a_e=6378.1\pm 0.1$ km and $GM/a_e=62.494\pm 0.01$ km^2/sec^2), θ is the geocentric latitude, which differs only negligibly from the geodetic latitude in terms other than J_2, and λ is the longitude. The values P_n^m are the associated Legendre functions defined by

$$P_n^m\sin\theta=\frac{\cos^m\theta}{2^n n!}\sum_{l=0}^{k}\frac{(2m-2l)!}{(n-m-2l)!}(-1)^l\sin^{n-m-2l}\theta, \tag{2}$$

where

$$\left.\begin{aligned} k&=\frac{n-m}{2} && (n-m \text{ even}),\\ k&=\frac{n-m-1}{2} && (n-m \text{ odd}).\end{aligned}\right\} \tag{3}$$

Prior to artificial satellites only J_2^0 was known and the value used at that time has been found to be inaccurate by 1 part in 300.

The equation of motion of a satellite can be written in terms of the earth's potential U. A principal problem of celestial mechanics is to determine the effect that the various J_n^m and K_n^m values have on the orbital elements. Six orbital elements are required to describe the Kepler ellipse with one focus at the origin: a, the semimajor axis; e, the eccentricity; i, the inclination of the orbit to the equator; Ω, the angular

TABLE 1.—*Estimates of Zonal Harmonics*

Source	$J_2^0 \times 10^6$	$J_3^0 \times 10^6$	$J_4^0 \times 10^6$	$J_5^0 \times 10^6$	$J_6^0 \times 10^6$	$J_7^0 \times 10^6$
O'Keefe, Eckels, and Squires (Reference 11)	1082.49 ± 0.06	-2.39 ± 0.26	-1.70 ± 0.06	-0.30 ± 0.53	–	–
Kozai (Reference 12)	1082.19 ± 0.02	-2.29 ± 0.02	-2.13 ± 0.04	-0.23 ± 0.02	–	–
King-Hele (Reference 13)	1082.79 ± 0.15	–	-1.4 ± 0.2	–	0.9 ± 0.8	–
Kaula (Reference 14)	1082.3 ± 0.2	-2.3 ± 0.1	-1.8 ± 0.2	-0.3 ± 0.2	–	–
King-Hele (Reference 15)	1082.7 ± 0.3	-2.4 ± 0.1	-1.7 ± 0.3	-0.1 ± 0.2	0.8 ± 0.1	-0.4 ± 0.2

coordinate of the node, or intersection of the orbit and the equatorial plane; ω, the argument of perigee or the angle from the ascending node to the point (perigee) of closest approach on the ellipse to the origin; and f, the true anomaly, the angle from perigee to the satellite.

In writing the disturbing potential in terms of the elements of a Keplerian orbit it is apparent that J_n^0, n even, will give rise to secular changes in ω and Ω but have no effect on a, e, or i. J_n^0, n odd, will cause periodic variations with frequency ω in e, i, ω, and Ω.

Since the first satellite launching, a large number of papers have been devoted to the reduction of orbital data to obtain estimates of J. Table 1 lists estimates of the zonal harmonics derived from studies of satellite motion. The tesseral and sectorial harmonics give rise to daily and semi-daily oscillations in the orbit. Using observations of the satellite Transit IV-A (1961o1), R. R. Newton obtained $J_{22}=2.2\times10^{-6}$ and $J_{41}=1.25\times10^{-6}$.[16]

The errors caused by neglecting higher order terms in the expansion are not yet known, and there is no proof that the accumulated effect of these terms is negligible. Since the series must converge the coefficients must decrease, but the rate at which they decrease is unknown. The high order harmonics may be of more or less the same magnitude. It seems unlikely that any given harmonic would be much larger than those already found, since this would demand an unexpected asymmetrical concentration of mass.

A major result of satellite studies is the discovery that the earth's gravitational potential is irregular. This reduces the value of the concept of flattening, since the earth's geometrical figure cannot be expressed in terms of a single quantity as earlier theories presumed. The surprisingly large magnitude of the low order harmonics clearly shows that the earth's figure is complex.

Figure of a Fluid Earth

A rotating fluid develops a bulge represented in the gravitational potential by finite values of J_n^0, n even. The difference between observed J_n^0 values for the earth and the J_n^0 values for an equivalent rotating fluid indicates the deviation of the earth from a perfect fluid. The moment of inertia of a fluid is related to the hydrostatic flattening f by:

$$\frac{C}{Ma_e^2} = \frac{2}{3}\left\{1 - \frac{2}{5}\left[\frac{5m}{2f}\left(1 - \frac{3}{2m}\right) - 1\right]^{\frac{1}{2}}\right\}, \quad (4)$$

to first order in the flattening, where m is the ratio of the centrifugal to the gravitational acceleration at the equator. With the observed value of C/Ma_e^2 the hydrostatic value of the flattening follows from the equation. Henriksen[17] and O'Keefe[18] noted that J_2^0 determines a much-improved estimate of the polar moment of inertia, since

$$\frac{J_2^0}{H} = \frac{\frac{C-A}{Ma_e^2}}{\frac{C-A}{C}} = \frac{C}{Ma_e^2}. \quad (5)$$

The hydrostatic flattening can then be used to obtain a value for the hydrostatic J_2^0:

$$J_2^0 = \frac{3}{2}f\left(1 - \frac{1}{2}f\right) - \frac{1}{3}m\left(1 - \frac{3}{2}m - \frac{2}{7}f\right). \quad (6)$$

This theory carried to higher orders, provides

TABLE 2.—*Comparison of Observed and Equilibrium Values of J_n^0 in parts per million*

Method	J_2^0	J_3^0	J_4^0	J_5^0	J_6^0	J_7^0
Observed	1082.7	-2.4	-1.7	-0.1	0.8	-0.4
Hydrostatic	1071.0	0	-2.9	0	0.00	0
Nonhydrostatic	11.7	-2.4	1.2	-0.1	0.8	-0.4
Topography and Perfect Isostasy	–	0.20	-0.33	0.69	-0.23	–

improved estimates of the hydrostatic values for J_2^0, J_4^0, and J_6^0.[18] Thus the observed value of J_2^0 can be used to obtain a description of what would be the gravitational potential of the earth if the earth behaved as a perfect fluid in hydrostatic equilibrium. Table 2 lists the hydrostatic values of J_n^0. It should be noted that the values of J_n^0, n odd, vanish for a fluid earth, as do the sectorial and tesseral harmonics.

The observed and hydrostatic values for the gravitational potential are compared in the first two lines of table 2. The third line gives the nonhydrostatic component of the potential computed as the difference between the observed J_n^0 and the J_n^0 for an equivalent rotating fluid. The nonhydrostatic J_n^0 represents density anomalies within the earth. The last line of the table lists values of J_n^0 arising from the continent-ocean structure provided the continents and oceans are in perfect isostatic equilibrium. The observed deviations from equilibrium are larger than and opposite in sign to those expected from the near-surface structure. These anomalies cannot be explained by near-surface or crustal variations in density.[14–19] The negative correlation between the continent-ocean structure and the gravitational figure was anticipated by Jeffreys on the basis of the surface observation of gravity.[20] But the magnitude of the deviations is surprisingly large. This negative correlation may have significance since it possibly reflects the process of formation of the continent-ocean system.

Strength of the Earth

The interpretation of the deviation of the earth's potential from that expected for an equivalent rotating fluid raises the problem of the anelastic response of the deep earth to small stresses. The question of the earth's response to large-scale small-amplitude stress differences imposed for long periods has plagued geophysicists for a century. Kelvin assumed that the earth can be treated as an elastic body even for long period deformations, whereas Darwin proposed that the earth behaves plastically. Today, proponents for both points of view can be found; the gravitational field as determined from the satellite orbits has yielded important information on this problem but no decision is yet possible.

The simplest interpretation of the density anomalies within the earth supposes that the anomalies are supported by the finite strength of the rocks making up the mantle. The anomalies lead to stress differences, the magnitude of these differences depending on the radial distances over which the stress differences are supported. If the entire mantle is involved, the differences between the observed and hydrostatic J_n^0 values imply maximum stress differences of the order of tens of bars. If only the upper mantle supports the density anomalies then higher stress differences are implied.

An alternative to the finite-strength interpretation for the density anomalies is that the values of J_n^0 indirectly result from convection currents; the density variations are directly associated with temperature inhomogeneities, the anomalies being supported by viscous stresses developed in the convective motion. Many geologists and geophysicists argue that the large scale structure of the earth can be interpreted only in terms of convective motion.[21] Perhaps the most striking evidence is that presented by Vacquier, Raff, and Warren who showed that great blocks of the Pacific Ocean floor off the California coast have

moved as much as 1000 km relative to each other.[22] This relative motion may be interpreted in terms of a crust being dragged along by subcrustal convective systems.

Prior to the acquisition of geodetic data from the analysis of satellite orbits the principal evidences for a mantle of zero strength were: (1) The figure of the earth closely approximates that of an equivalent rotating fluid; (2) Formerly glaciated areas in Scandinavia and North America have been rising—this was interpreted as a viscous rebound following the removal of ice.

The presatellite value used for the flattening of the earth, about 1/297.2, agrees closely with Bullard's calculation of the flattening of a rotating fluid having the estimated density distribution of the earth.[23] This close agreement indicated that the earth behaves as a fluid for long term deformations. The new data obtained from satellites indicate that the actual bulge of the earth is about $\frac{1}{2}$ percent larger than the surface of an equivalent rotating fluid. Munk and MacDonald show that the bulge is to be expected, in a sense, for a rotationally decelerating earth.[19,24] At the present rate of deceleration, the bulge implies an equilibrium figure about 10^7 years ago. The adjustment of the bulge to the changing rotation rate can result from continuous creep of the mantle at the stress difference fixed by the creep strength, or from intermittent flow or fracture as the strength is exceeded locally. If the entire mantle is undergoing more or less continuous adjustment, the discrepancy between the actual and equilibrium bulge of the earth gives a good estimate of the creep strength of the mantle, about ten bars.

The classical estimate of the viscosity of the mantle depends on the interpretation of uplift in certain Pre-Cambrian shield areas. During the last glacial epoch, the Baltic and Canadian shields were depressed by the weight of ice; and isostatic uplift followed the removal of this glacial ice. Haskell solved the problem of loading and unloading of a layer of viscous fluid and determined the viscosity of the mantle from the rate of uplift.[25,26] His value of 3×10^{21} in cgs units depends critically on the assumption that the uplift is due to the viscous response of a material possessing zero strength. This assumption has been severely criticized, particularly by Russian investigators.[27] Gravity studies on the formerly glaciated regions reveal a close correlation of negative anomalies with uplift, in agreement with the hypothesis that the area is undergoing isostatic uplift. But there are large areas in Norway where positive anomalies are found in regions undergoing uplift. Furthermore, the uplift in Scandinavia is of approximately the same magnitude as the uplift in other shield areas that have not undergone glaciation. Lyustikh suggests that the uplift is of a general character and that the association with glaciation is accidental. A further difficulty in accepting a viscosity of the order proposed by Haskell is the short time scale implied for vertical motion. Large scale gravity anomalies should persist for longer than 10^4–10^5 years if the crust rested on a fluid with a viscosity of 10^{22} in cgs units. This is in direct contradiction to the existence of ancient geologic features associated with large gravity anomalies.

Section 3

BODILY TIDES AND THE ROTATION OF THE EARTH

The earth's gravitational potential provides information regarding its internal density distribution; and the difference between the observed potential and the potential calculated for an equivalent rotating fluid yields data on the earth's anelastic behavior. Because of the large scales involved, it is assumed that the long term anelastic responses are associated with the gravitational anomalies. As we shall see in the next section, data from seismology provide information on the anelastic response of the earth to oscillations with periods on the order of hours to seconds. The earth also yields to other distorting forces having periods intermediate between the very long secular changes and the high frequency seismic deformations.

The attraction of the moon and the sun produces deformations in the body of the earth analogous to the tides in the oceans. The attraction of an external body on a spherical earth may be considered to be of two parts. First, there is a net force attracting the earth toward the external body in question which is of such a magnitude that it would seem that all the earth's mass were

concentrated at its center of gravity. This force is responsible for keeping the earth in its orbit around the sun or around the center of gravity of the earth-moon system. Second, there is a differential force on the various parts of the earth's volume because they are at slightly different angular positions from the center of the sun or moon. These differential or tide-raising forces produce the ocean tides. They also deform the solid earth, tending to elongate it into an ellipsoid whose major axis is directed toward the disturbing body. The ocean tides are large and more obvious, but the deformation of the solid earth also produces observable effects which can be measured. Therefore, these effects yield information on the interior of the earth. The "earth tides" yield readily to a mathematical treatment, whereas the ocean tides do not. The free periods of elastic vibrations in the earth are on the order of an hour and an equilibrium theory is therefore applicable to the slowly varying tidal deformations. In the oceans the free periods are probably on the order of days, so they cannot be treated as equilibrium phenomena. As we shall see, ocean tides complicate the analysis of body tides.

In addition to the forced motion induced by the action of the sun and moon, there is a free nutation of the earth's axis of rotation. Euler showed that a body symmetrical about its axis of rotation would rotate permanently about the axis of greatest moment. A small disturbance makes the axis of rotation move in a cone about the axis of greatest moment and it completes its revolution within the body in a period of $A/(C-A)$ times the period of rotation. The ratio $A/(C-A)$ is well known, from the precession of the equinoxes, to be 305. The wobble in the pole of rotation should show up in astronomical observations as a change in latitude with about a 10 month period. During the nineteenth century a number of astronomers searched for changes in latitude with a 10 month period. The results were inconclusive. Chandler[28] announced a term in the latitude variation with a period of 428 days, 40 percent longer than Euler's classic value; in addition, Chandler detected a 12 month term. The discovery of the 428 day term was entirely unexpected and raised doubts concerning the validity of the observations. But only a year later Newcomb was able to demonstrate that the yielding of the earth and its oceans could bring about an increase in the period from 10 to 14 months. He attributed $\frac{1}{4}$ of the increase to the mobility of the oceans, the remainder to the earth's anelasticity. The lengthening of the period can be understood to occur as follows: The rotation of the earth produces an elastic strain which is symmetrical about the instantaneous axis and does nothing to displace it. This is superimposed on a permanent flattening, the two together making up the total flattening. The permanent part affects the axis of rotation just as such a condition would affect a rigid body, and it is this part that determines the period; but both parts are attracted by the sun and moon and contribute to the precession. The lengthening of the period from 10 months to 14 months thus gives a measure of the elastic yielding of the earth. In addition, the Chandler nutation or wobble is not constant in time but decays. If a source is given, the rate of decay provides information regarding the anelasticity of the earth.

Love Numbers

A description of the deformation of the earth due both to the tide-raising forces of the sun and moon and to the 14 month wobble is best given in terms of certain numbers introduced by Love. Consider the earth's response to a disturbing potential $U(r)$ described in terms of a spherical harmonic of degree 2; the tidal forces of the moon and the sun and centrifugal forces arising from the earth's rotation can be written as gradients of such a potential U. The resulting deformation defines the Love numbers h and k as follows:

The ground is lifted by an amount hU/g, where g is the surface value for the gravitational acceleration. The additional gravitational potential at the displaced surface arising solely from this redistribution of mass is kU. The numbers h and k are the Love numbers of degree 2. Thus $1+k$ is a factor allowing for the attraction of the bulge by itself and the response of hU/g takes this self-attraction into account. A fluid surface covering the globe would remain equipotential and be lifted by $(1+k)U/g$ relative to the center of the earth and by $(1+k-h)U/g$ relative to the sea bottom.

In addition to the vertical displacement of the solid surface by hU/g there is horizontal displacement with the components

$$\left.\begin{aligned} &\frac{l}{g}\frac{\partial U}{\partial \theta}, \\ &\frac{l}{g}\frac{1}{\cos\theta}\frac{\partial U}{\partial \lambda}, \end{aligned}\right\} \qquad (7)$$

where the third Love number l describes the horizontal yielding of the surface.

The Love numbers are dimensionless parameters which neatly summarize some of the earth's elastic properties. Evaluation of them is somewhat difficult, and we will consider several sources of information in our discussion of them.

Secular Love Numbers

Consider the distortion of the earth due to any potential U of degree 2. The distortion gives rise to an exterior gravitational potential

$$k\frac{a^5}{r^5}U \quad . \qquad (8)$$

The gravitational potential near the boundary of a nearly spherical body is given by MacCullagh's formula.[29] In the present case, the deformation is a spherical harmonic of degree 2 and the pertinent terms in MacCullagh's formula can be written as:

$$\left.\begin{aligned} U &= \frac{GM}{r} + V, \\ V &= \frac{G}{2r^5}\left[C_{11}(x_2^2 + x_3^3 - 2x_1^2) + \cdots - 6C_{12}x_1x_2 - \cdots\right] = k\frac{a^5}{r^5}U, \end{aligned}\right\} \qquad (9)$$

where C_{ij} are the components of the moment of inertia tensor. The dots in Equations 9 indicate two additional terms that are obtained by the cyclic rotation of subscripts.

Now consider the special case of centrifugal potential,

$$\frac{1}{2}\left[\omega^2 r^2 - (\omega_i x_i)^2\right], \qquad (10)$$

where $\omega^2 = \omega_i\omega_i$ and $r^2 = x_ix_i$. Expression 10 can be organized into terms of the form

$$\frac{1}{3}\omega^2 r^2 + U \quad , \qquad (11)$$

where

$$U = \frac{1}{6}\left[\omega_1^2(x_2^2 + x_3^2 - 2x_1^2) + \cdots - 6\omega_1\omega_2 x_1x_2 - \cdots\right] \qquad (12)$$

is a spherical harmonic of degree 2. The term $\frac{1}{3}\omega^2r^2$ leads to a purely radial deformation which consists of a contraction near the center of the earth and extension in the outer parts. Combining equation 12 with equations 9 yields

$$C_{ij} = I\delta_{ij} + \frac{ka^5}{3G}\omega_i\omega_j + \text{constant} \quad , \qquad (13)$$

where

$$I = \frac{1}{3}(C_{11} + C_{22} + C_{33}) \qquad (14)$$

is the inertia of the sphere in the absence of rotational deformation; and δ_{ij} vanishes except for $i=j$, when it is unity. This determines the constant so that, finally,

$$C_{ij} = I\delta_{ij} + \frac{ka^5}{3G}\left(\omega_i\omega_j - \frac{1}{3}\omega^2\delta_{ij}\right) \quad . \qquad (15)$$

The Love number k in equation 15 can be interpreted as a measure of the earth's yield to centrifugal deformation during the last 5 million years.

Without loss of generality, the x_3 axis can be placed along the rotation vector. Then $\omega_1=0$, $\omega_2=0$, $\omega_3=\Omega$, and

$$\left.\begin{aligned} C_{11} = C_{22} = A &= I - \frac{k_s a^5}{9G}\Omega^2 , \\ C_{33} = C &= I + \frac{2k_s a^5}{9G}\Omega^2 , \\ k_s &= \frac{3GHC}{a^5\Omega^2} . \end{aligned}\right\} \quad (16)$$

For a homogeneous sphere, $C=\frac{2}{5}Ma^2$; then with $M=5.98\times10^{27}$ gm for the mass of the earth, $k_s=1.14$. In reducing the value of C from the satellite observations, the value for k_s becomes

$$k_s=0.96. \quad (17)$$

The foregoing calculation of k_s involves the observed rate of precession and the gravitational potential of the earth derived from the observations of satellites. No assumptions were made concerning the stress-strain relations within the earth.

The secular Love number measuring the response of the earth to very-long-period stresses is far different from the Love numbers obtained from studying the earth tides and the Chandler wobble. There can be no doubt that the earth responds far differently to tidal forces than it does to diurnal rotation.

Tidal Variation of Gravity

Consider an external disturbing function U_2. The total external gravitational potential is

$$U = \frac{ga^2}{r} + U_2 + \frac{ka^5}{r^5}U_2 . \quad (18)$$

The tidal deformation displaces the surface and to first order the acceleration at $r=a$ is

$$g\left[1 - \frac{2U_2}{ag}\left(1 - \frac{3}{2}k + h\right)\right] . \quad (19)$$

Thus, gravity measured at the surface of the earth will show a tidal variation. The effect of the yielding of the earth is to cause a larger variation of gravity than would otherwise be anticipated.

The tidal periods are very much longer than the longest free period of oscillation (see the next chapter), so the yielding should be nearly in phase with the tidal driving forces. The departure from perfectly elastic behavior, such as might result from energy dissipation due to frictional processes within the earth, produces a phase lag in the yielding and, hence, in its contribution to the total gravity variation.

Observations of the variation of gravity with a sensitive gravity meter provide data on the amplification factor $(1-\frac{3}{2}k+h)$ and thus on the possible values of the Love numbers h and k. Tidal variations in gravity have been determined at a number of stations distributed over the earth's surface. Of particular interest are measurements carried out by the Institute of Geo-

TABLE 3.—*Amplification of Diurnal and Semidiurnal Tides by Elastic Yielding of the Earth (from unpublished data of Harrison, Ness, and Slichter)*

Tidal Component	Period (hours)	$1-\frac{3}{2}k+h$	Phase Lag (degrees)
O_1	25.84	1.177 ± 0.018	+0.64 ± 1.07
P_1 K_1	23.92	1.140 ± 0.012	+0.26 ± 1.85
M_2	12.42	1.1821 ± 0.007	+0.86 ± 0.74
S_2 K_2	12.00	1.1849 ± 0.0072	+0.12 ± 0.72

physics of the University of California with the use of a LaCoste–Romberg gravimeter. The principal tides for which observations have been made are those having a period near 12 and 24 hours. Table 3 lists results obtained by the Institute during the International Geophysical Year from the measurements of twelve stations distributed around the earth. The determinations are based on observations of about 6 weeks duration at each station. Uncertainties are due primarily to the perturbing influence of ocean tides. Ocean tides, unlike body tides, are not an equilibrium phenomenon, and since they differ in phase from the disturbing potential the local value of acceleration is perturbed.

Chandler Wobble

Observations of the latitude of stations belonging to the International Latitude Service determine the wobble of the instantaneous axis of rotation. A spectral analysis of the variation of latitude shows two principal peaks, one with a period of a year and another with a period of 429 days.[19] The lengthening of the period over that of a Eulerian rigid body is due both to the elastic yield of the earth and to the fluid of the core and ocean.

The period of wobble can be interpreted in terms of a tidal effective Love number $k=0.29\pm 0.01$. In interpreting this value it is important to realize that the earth as a whole is involved, including both the oceans and the fluid inner core. The oscillations excited by the wobble are only in part communicated to a liquid core so that the inertia of the core plays only a minor role in determining the period. Indeed, if the core were a perfect fluid occupying a spherical cavity in a rigid shell, the core would not partake in the wobble at all. The fraction of the core involved in the wobble depends on the ellipticity, density distribution, and viscosity of the core, and on the electromagnetic forces acting on the core and mantle. Only the variations in ellipticity and density due to the finite elasticity of the shell have been investigated. The subtraction of the inertia of the core reduces the period of the Chandler wobble, as compared to a static model in which the earth is taken to be an elastic body. The reduction is in the ratio of the inertia of the core to the inertia of the core plus the shell. For a rigid outer earth the ratio is $\frac{1}{10}$ and the reduction in the period is 30 days.

In quite the same way an inviscid ocean covering the whole earth would shorten the period, though only by a very small amount since the ocean's moment of inertia would be minute. If, on the other hand, this ocean remained always aligned with respect to the wobbling axis, then the ocean would lengthen the period. The lengthening due to such an equilibrium tide would be 33 days. If, without any real evidence, we assumed an equilibrium pole tide, then the value of the effective Love number would change from 0.29 to 0.23. Jeffreys and Vincente showed that the effects of the core increased k by about 0.08.[30] The Love number due to the elastic deformation of the earth would then be $k=0.31$.

The observed Chandler wobble frequency of 0.85 cycles per year (cpy) can be interpreted as the earth's Eulerian frequency of nutation (1.20 cpy) diminished by about 0.25 cpy because of the elastic yield of the earth and by about 0.1 cpy because of the fluid yield of the oceans. The last value is very uncertain since it is questionable whether the response of the ocean is an equilibrium response at these low frequencies.

In addition to the frequency of the Chandler wobble and its interpretation in terms of the elastic properties, it is of interest to examine the problem of dissipation and generation of the wobble. Dissipation is important since the anelastic response of the elastic mantle to a 14 month excitation is involved, though the interpretation involves subtracting the contributions of the core and oceans to the dissipation. The Chandler wobble may be regarded as a resonance amplification of some unknown excitation. In this sense, it is like playing a relatively broadband noise through a filter with a fairly narrow passband. The relative bandwidth of the filter, Q^{-1}, can be obtained from the latitude observations. The resulting value depends on what series of data are used and the value ranges from 18 to 60.

First we need to estimate the rate of energy dissipation. The kinetic energy of a rigid body

freely wobbling about a mean axis, x_3, is

$$K = \frac{1}{2}\left(A\omega_1^2 + A\omega_2^2 + C\omega_3^2\right) . \quad (20)$$

The kinetic energy of the rigid body in steady rotation having the same angular momentum is

$$K_0 = \frac{1}{2C}\left[A^2\left(\omega_1^2 + \omega_2^2\right) + C^2\omega_3^2\right] . \quad (21)$$

The time variance of the kinetic energy is then

$$\Delta K = K - K_0$$

$$= \frac{1}{2} HA\Omega^2\left(m_1^2 + m_2^2\right) , \quad (22)$$

where

$$m_i = \frac{\omega_i}{\Omega} . \quad (23)$$

For a root-mean-square amplitude of 0.14 seconds of arc for the Chandler wobble,

$$2\Delta K = Q \times 10^{21} \text{ ergs}$$

$$= 4 \times 10^{-15} K \quad (24)$$

is the mean kinetic energy of the Chandler wobble; and

$$2\Delta K \sigma_0 Q^{-1} = 10^{15} Q^{-1} \frac{\text{ergs}}{\text{sec}} , \quad (25)$$

(where σ_0 is the frequency of the wobble) is the mean rate of energy dissipation. This rate is very small compared to the energy dissipated by tides and to the energy loss due to radioactivity.

Interpretation of the damping of the wobble has had a long and complex history. It was first attributed by Jeffreys to tidal friction in the oceans. Bondi and Gold dismissed the core and oceans as a possible sink, since neither has a sufficient moment of inertia to affect the mantle.[31] However, Munk and MacDonald argue that either the core or mantle is sufficient to account for the damping and that the low Q observed for a Chandler wobble does not represent the anelastic response of the mantle.

The excitation of the Chandler wobble remains a problem. Early researchers thought that the wobble was due to variations in mass distribution on the earth's surface. These variations would produce time-dependent products of inertia which in turn would result in a shift of position of the axis of rotation. However, a detailed study by Munk and Hassan shows that the computed spectral density of the atmospheric variation at the Chandler frequency falls short of the energy required in the wobble by one or two orders of magnitude[32] MacDonald (unpublished data) has repeated the analysis of the meteorological forcing function, using somewhat improved methods of data reduction. The results of Munk and Hassan have been confirmed. However, a bispectral analysis shows that the energy at one cycle per year interacts nonlinearly with energy at 0.15 cycle per year. The excitation of the Chandler wobble is due to nonlinear coupling. Deviations from perfect elasticity or magnetohydrodynamic coupling of the core and mantle can provide the nonlinearity. The latter possibility is consistent with present theories regarding the core-mantle interaction.[42]

Values for the Elastic Love Numbers

The calculation of the Love numbers for an elastic earth is a straightforward problem in elastic theory. The methods are similar to those used in treating the earth's free oscillations and will be considered in the following chapter. Takeuchi, Saito, and Kobayashi[33] have determined the Love numbers for a model of the earth having a velocity distribution given by Gutenberg (table 4).

The Love number k calculated for an elastic earth can be compared with k for the Chandler wobble. They are in close agreement, but this agreement is fictitious since the Love number for the Chandler wobble involves the response of the oceans and fluid core, neither of which are considered in the calculation of the elastic Love number. The gravitational amplification $(1-\frac{3}{2}k+h)$ for an elastic earth is 1.172. This value can be compared with the values listed in table 3. The differences do not imply uncertainties in the elastic constitution of the earth, but rather, the nonequilibrium response of the ocean tides complicates the interpretation of the semidiurnal and diurnal tides.

From the study of earth tides and the Chandler

TABLE 4.—*Love Numbers for an Elastic Earth for the Gutenberg Model*

n	k_n	h_n	l_n
2	0.280	0.592	0.076
3	0.083	0.274	0.010
4	0.035	0.161	0.007
5	0.020	0.116	0.006
6	0.013	0.094	0.005
7	0.009	0.081	0.004
8	0.007	0.073	0.003
10	0.005	0.063	0.002
12	0.003	0.055	0.002
14	0.003	0.051	0.001
16	0.002	0.048	0.001

wobble, we find $h=0.59$ and $k=0.29$. These are in marked contrast with the values of the *secular* Love numbers $h_s=1.96$, $k_s=0.96$ obtained from the figure of the earth. A number of hypotheses have been suggested to explain the difference and it is not known which, if any, is correct. One hypothesis is based on the relative magnitude of stresses, with the secular Love numbers referring to stress differences above the critical strength, and tidal effective Love numbers to stress differences below the critical strength. A second hypothesis is based on the relative duration of stresses. In this case the secular Love numbers refer to stresses much more prolonged than some critical duration and the tidal effective Love numbers refer to stresses much less prolonged than this critical duration. Still another possibility is that the earth was originally molten and now has the hydrostatic figure appropriate at the time of congealment. This last hypothesis is extremely unlikely, both on the grounds of the thermal history of the earth and also because of the observed deceleration of the earth due to the lunar tide.

Load Love Numbers

In many geophysical problems it is convenient to express the effect of surface load in terms of dimensionless numbers analogous to the Love numbers. Consider the effect of a variable surface slope, $q(t)$ gm/cm^2. We are concerned with the term q_n of degree n in the surface spherical harmonic expansion of q. The interior potential resulting from a gravitating layer q_n is

$$U_n = \frac{4\pi G a q_n}{2n+1}\left(\frac{r}{a}\right)^n. \qquad (26)$$

TABLE 5.—*Load Love Numbers*

n	h_n'	k_n'
2	-1.034	-0.312
3	-1.078	-0.191
4	-1.083	-0.126
5	-1.121	-0.096
6	-1.185	-0.081
7	-1.260	-0.072
8	-1.338	-0.066
10	-1.486	-0.058
12	-1.622	-0.052
14	-1.750	-0.048
16	-1.872	-0.045

Deformation is due to two opposing effects: a normal stress due to the load, which will depress the surface; and a gravitational attraction of the earth on the load, which will raise the ground. The combined effect of pressure and attraction raises the surface by $h_n'U_n/g$ and the gravitational potential arising from the distortion is $k_n'u_n'$. This defines h_n' and k_n'. The depression is found to be somewhat larger than the gravitational uplift and the load Love numbers are negative. The calculated values for the load Love numbers are shown in table 5.

Section 4

ELASTICITY OF THE EARTH AT HIGH FREQUENCIES

In modern analysis the variation, with depth, of the two seismic body waves (compressional and shear) is used to determine the internal constitution of the earth. Velocity-depth curves may be derived from distance-time curves; the latter are constructed by identifying, interpreting, and analyzing the various phases of the seismogram.[34] The variable quality of the records, the local effects, and the regional grouping of stations all complicate the analysis. Moreover, the conversion of a distance-time curve to a velocity-depth curve depends upon the mathematical methods applied.

The observation analysis of the earth's free oscillations provides another method for determining the general features of the velocity-depth curves. Since the oscillations involve the earth as a whole, difficulties associated with analyzing many events on different seismographs for several earthquakes should vanish. The results from the Chilean earthquake of May 22, 1960, show that Gutenberg's solution for the variation of velocity with depth in the mantle fits the observations very well (figure 1). Before discussing the earth's free oscillations, it would be well to consider the features of the velocity-depth curves on which most seismologists agree.

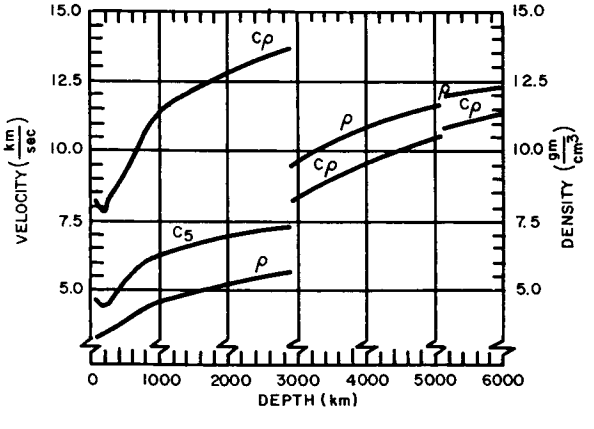


FIGURE 1.—Variation of the compressional wave velocity C_p, the shear wave velocity C_s, and the density ρ within the earth (reference 35).

Under continental areas the velocity in the upper few tenths of a kilometer varies markedly with depth and location. The velocities, C_p, of the compressional wave range from 5 to 7 km/sec. At a depth averaging some 30–35 km there is a discontinuity in the seismic velocity, and below this discontinuity the velocity is usually above about 8 km/sec. The material above this discontinuity, first recognized by Mohovovicic, is termed the crust. The crust is much thinner under oceans, with some 6 km of water underlain by some 4 km of rock. At this depth (10 km) the velocities undergo a transition. The region below the crust, termed the mantle, extends to a depth of 2898 km. In the Gutenberg solution for the variation of velocity with depth, the velocity initially decreases, reaching a minimum at a depth of 150 km; then it rapidly increases, to a depth of 900 km. Below this it increases more gradually. The existence of two velocities throughout the mantle is significant and suggests a normal type of elasticity. The absence of shear waves below the mantle is the principal reason for believing this region to be liquid. The absence of this velocity and the drastic decrease of the compressional wave velocity are the outstanding seismic features which distinguish the core from the mantle. The surface of separation of the core and mantle is usually considered to be sharp and is the best located, as well as the largest, discontinuity in the earth's interior. At a depth of 5121 km another discontinuity is encountered. The velocity distribution within the inner core is not known.

The Earth's Free Oscillations

During the great Chilean earthquake of May 22, 1960, the earth's free modes of vibration were

observed for the first time. Their detection broadened the spectrum over which a geophysicist may study the earth's interior. Prior to 1960 almost all information regarding the earth's interior had been derived from detailed investigations of the arrival of the elastic body waves as recorded by seismographs. These waves travel different paths through the earth and contain most of the energy in the high frequency part of the spectrum (10–0.1 cps). The interpretation of arrival times is based on a ray theory similar to that of geometrical optics.[34] The use of low frequency normal modes as a tool for investigating the earth is somewhat analogous to the astronomers' use of radio frequencies as a supplement to observations in the visual range. In the following sections we will consider the general features of the earth's free oscillations. The detailed calculations are contained in Appendix A.

The Free Vibrations of an Elastic Sphere

A small-amplitude disturbance in an elastic solid is carried outward by two traveling waves. The fast (compressional) wave carries the compression and rarefaction of ordinary sound at a phase velocity of

$$C_p^2 = \frac{\lambda + 2\mu}{\rho} , \qquad (27)$$

where λ and μ are the Lamé constants and ρ is again the density. The slow (shear) wave transmits particle motion at right angles to the direction of propagation with a phase velocity of

$$C_s^2 = \frac{\mu}{\rho} . \qquad (28)$$

The adiabatic bulk modulus k is given by

$$\frac{k}{\rho} = C_p^2 - \frac{4}{3} C_s^2 . \qquad (29)$$

If an elastic solid is sufficiently isolated from its surroundings, reflections from boundaries may set up standing waves. Lamb showed that the vibrations of an elastic sphere can be classified into two groups: In the *toroidal* or torsional oscillations, a particle executes motion on the spherical surface; there is no radial component of motion. The toroidal oscillations unite to form the familiar horizontally polarized shear waves of classical seismology. The *spheroidal* oscillations combine both radial and tangential motions to produce compression and rarefaction. A degenerate spheroidal oscillation involves only radial motion; the entire sphere expands or contracts.

The notation adopted to describe the earth's free oscillations is similar in many ways to the notation used in spectroscopy. This reflects the common mathematical structure of these two fields. The solution to the equations of motion of an elastic sphere can be separated into a function dependent upon the angular coordinates and a function dependent upon radius. The angular function is written as a sum of spherical harmonics,

$$X_l^m = P_l^m(\sin\theta) e^{im\lambda} , \qquad (30)$$

where P_l^m is the associated Legendre function, θ is the latitude, λ is the longitude, and m and l are the magnetic and azimuthal quantum numbers, respectively. The time dependence can be included in the exponential:

$$\exp\left[i(m\lambda - \omega t)\right] , \qquad (31)$$

where ω is the angular frequency. This represents a wave traveling with a speed of m/ω radians per second. If m is positive the wave travels from west to east; if m is negative it travels from east to west. The two signs of m are symmetrical in a stationary sphere; however, rotation destroys this symmetry and creates important differences between waves traveling from west to east and those traveling from east to west. The values m and l must assume integral values because of the conditions on the outer boundary and the conditions for regularity at the center of the sphere. These integral values determine the surface pattern of deformation associated with a particular free oscillation. The number of lines of vanishing displacement associated with the angular coordinate θ is $l-m$; the number of nodal lines associated with the angular coordinate λ is m. There will also be surfaces of zero particle displacement associated with the radial function. The description of the radial function is in terms of n, the radial quantum number.

The free oscillations can thus be characterized by three values: l and m determine the pattern of displacement on the spherical surface and n

determines the number of internal nodal surfaces. The notation that has been adopted is ${}_nS_l^m$ and ${}_nT_l^m$, for spheroidal and toroidal oscillations, respectively. The expression ${}_nT_l^m$ denotes a toroidal oscillation with n radial nodal surfaces and a displacement pattern on the surface of the sphere fixed by the spherical harmonic with ordinal numbers l and m. In the ${}_0S_2$ oscillation a sphere alternately assumes a prolate and oblate form; this is sometimes termed the "football" mode. In the ${}_0T_2$ oscillation one hemisphere differentially rotates or twists relative to the other.

The eigen frequencies for an inhomogeneous spherical shell can be determined by numerical computation (Appendix A). The distribution of elasticity and density is taken to be that of the Gutenberg model of the earth (figure 1). The effects of the inhomogeneous elasticity and density on the earth's free oscillations are illustrated in figures 2 and 3 where the Gutenberg earth model is contrasted to a homogeneous model in which

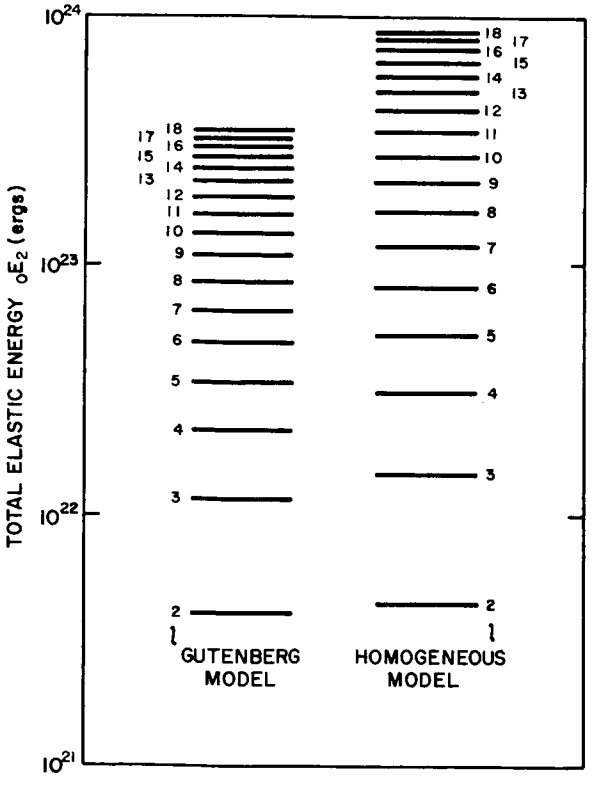


FIGURE 2.—Total elastic energy in the toroidal oscillation (${}_0T_l$). The energy is normalized to a 1 cm surface displacement.

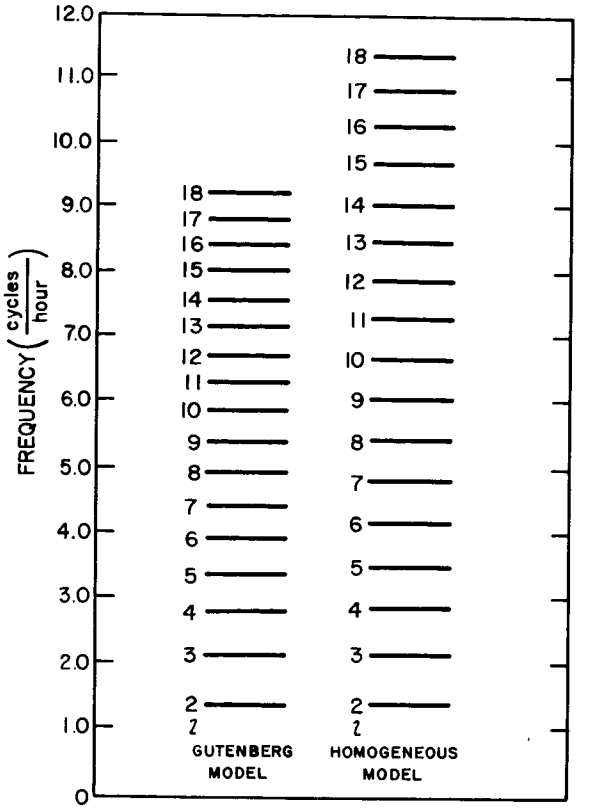


FIGURE 3.—Comparison of the resonant frequencies for the toroidal oscillations (${}_0T_l$) in the Gutenberg and homogeneous models.

the elasticity equals the average elasticity of the Gutenberg model. The energy levels shown in figure 2 are normalized to provide a 1 centimeter displacement at the surface (the observed displacement is much less, on the order of 10^{-6} cm). About 3×10^{21} ergs are required to produce this displacement with a ${}_0T_2$ surface pattern in both the homogeneous and inhomogeneous models. At higher modes and frequencies (figure 3) it takes more energy to form the complicated surface pattern of displacement and still maintain this maximum surface amplitude of 1 centimeter. The needed energy is greater in the homogeneous model since the near-surface rigidity is larger. The difference increases with greater values of the mode number l.

The elastic energy is more or less evenly distributed over the entire mantle at low mode numbers (figure 4). At higher mode numbers it is concentrated in the outer layers of the mantle.

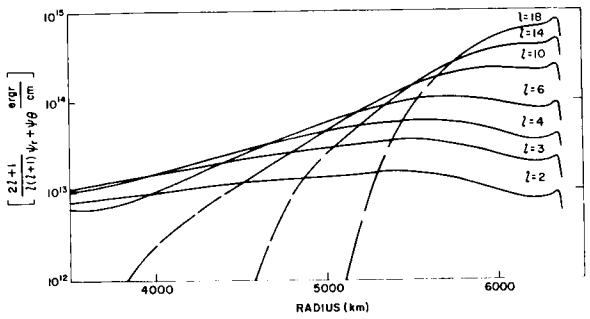


FIGURE 4.—Elastic energy per unit radius in the toroidal oscillations for the Gutenberg earth model.

The ${}_0T_2$ oscillation involves the mantle, but the ${}_0T_{18}$ oscillation is confined to the upper few hundred km.

Observations of the Earth's Free Oscillations

Despite considerable theoretical efforts the first attempt to observe the earth's oscillations was made only recently. Benioff constructed a strain-measuring seismometer, a silica glass rod 24 meters long, to investigate the low frequency spectrum. He suggested that an apparent 57 minute periodicity visible on the strain records of the Kamchatka earthquake of 1952 was the ${}_0S_2$ mode. This suggestion prompted extensive computational work. A second attempt at detecting the free oscillations was made in 1958 by spectral analysis of the background noise in the strain seismometer and measurement of the changes of the local gravitational field. The results were negative.

Several instrumental developments during 1958 and 1960 made possible the observation of the free oscillations excited by the Chilean earthquake of May 22, 1960. Benioff modified the circuitry associated with the seismometer so that the effect of the finite amplitude of the earth tides was reduced and a greater magnification achieved. A lower noise level was achieved on the LaCoste–Romberg gravimeter at the University of California. In addition, the Lamont Geological Observatory installed a strain gauge of the Benioff type in a mine shaft near Ogdensburg,

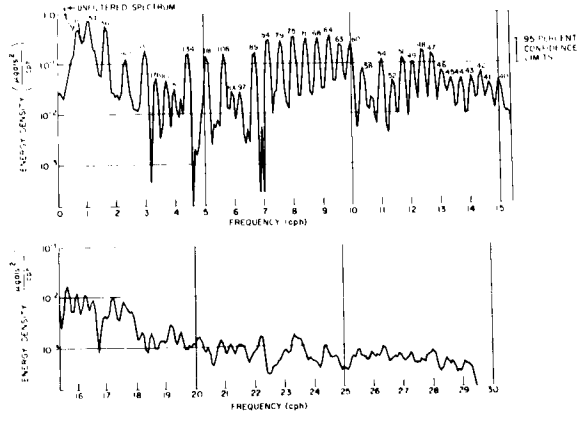


FIGURE 5.—Power spectrum of the gravity record after the Chilean earthquake of May 22, 1960, made with the University of California (Los Angeles) earth tide gravimeter; recording period, 110 hours, May 23–27; interval, 1 minute; sensitivity, 0.1 microgal.

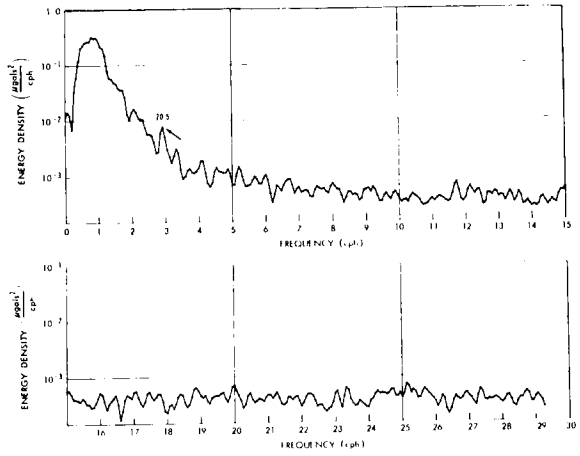


FIGURE 6.—Power spectrum of a quiet period 1 month after the Chilean earthquake made with the University of California (Los Angeles) earth tide gravimeter; recording period, 116 hours, June 23–28; interval, 1 minute; sensitivity, 0.1 microgal.

New Jersey. The strain seismometer and the gravimeter complement each other. The strain seismometer is sensitive to strain produced by both vertical and horizontal motion; it, therefore, records both spheroidal and toroidal oscillations. On the other hand, the gravimeter records only vertical accelerations and spheroidal oscillations. A combination of observations from the two instruments permits a separation and identification of vertical and horizontal motion. The free oscillations excited by the Chilean earthquake were detected on both instruments.

The power spectrum of the gravity variations in Los Angeles for 4 days following the Chilean earthquake is shown in figure 5. This figure should be compared with figure 6, a record of a quiet interval of 116 hours, 1 month after the earthquake, which is almost structureless except for a significant peak at 20.5 minutes.

Thus the spectrum of a seismic disturbance is characterized by well-defined short peaks for periods between about 8 minutes and 1 hour. At higher frequencies the isolated peaks begin to merge into a continuum as a result of the finite width of the individual peaks and the increased number of peaks. Similar analyses were carried out on the records of strain by Benioff, Press, and Smith[36] and by Alsop, Sutton, and Ewing.[37]

Comparison between Theory and Observation

A comparison between the spheroidal modes calculated by Pekeris, Alterman, and Jarosch[38] and those measured by Ness, Harrison, and Slichter[39] is shown in figure 7. At low frequencies the observations favor neither model, since only the average properties of the earth are involved;

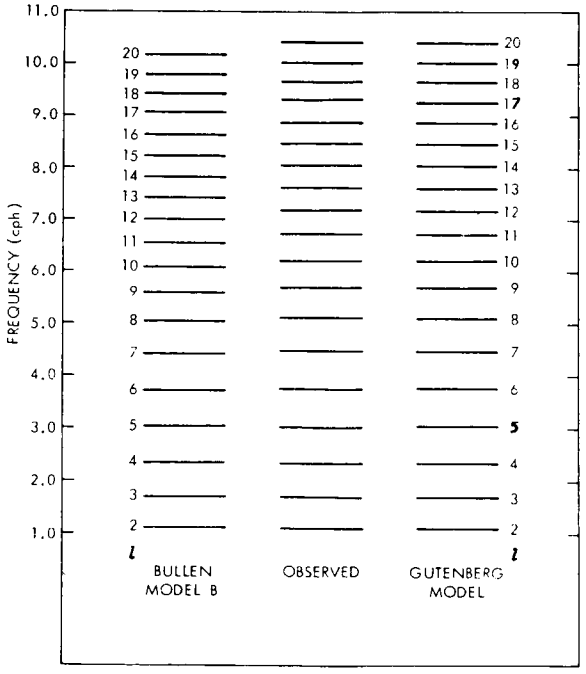


FIGURE 7.—Comparison of the calculated and observed frequencies of the spheroidal oscillations ${}_0S_l^0$.

TABLE 6.—*Comparison of Observed and Calculated Toroidal Frequencies (reference 40)*

Order	Bullen B	Gutenberg	Observed
2	44.2	43.6	42.9
3	28.6	28.2	28.6
4	21.9	21.6	21.9
5	18.1	17.9	18.0
6	15.6	15.4	15.5
7	13.7	13.6	13.6
8	12.3	12.2	12.3
9	11.2	11.2	11.2
10	10.3	10.3	10.3
11	9.59	9.66	9.61
12	8.95	8.94	9.06

but at higher frequencies the observations closely fit the Gutenberg model. A similar comparison can be made with the toroidal oscillations, and again (Table 6) the fit is better with the Gutenberg model.

Line Structure

The detailed reduction of the Chilean earthquake records shows that the low frequency spectral peaks appear as multiples instead of as the single lines that would be expected for a stationary elastic sphere. Rotation destroys the symmetry with respect to the integer m, and a degeneracy associated with the symmetry is removed. The effect of rotation on the oscillations can be qualitatively understood by recalling that a free oscillation is composed of a number of running waves. Waves traveling in the direction of the earth's rotation are carried forward relative to waves traveling in the opposite direction. The net effect is that the total pattern of surface deformation rotates relative to the earth. The local effect is to cause the vibrating particles to precess much in the manner of the Foucault pendulum. The rotational splitting is analogous to the Zeeman effect of spectroscopy, where a magnetic field removes the degeneracy with respect to the quantum number m. Detailed calculations of the splitting in the spheroidal oscillations have been made by Backus and Gilbert,[41] Pekeris, Alterman, and Jarosch,[38] and MacDonald and Ness.[42] Calculated splitting is in agreement

with the splitting observed in the low order spheroidal oscillations, an oscillation of order l being split into $2l+1$ peaks. The fine structure of the lowest order toroidal oscillations is in doubt and the line ${}_0T_2$ presents a number of problems.

If the earth were a perfectly elastic body, then the spectral peaks should show up as individual lines broadened only by the data reduction techniques (instrumental broadening). The deviations from perfect elasticity and fluidity result in a natural broadening of the lines. The degree to which a given line is broadened, or alternatively the rate at which a given peak decays, provides, in time, a measure of the anelastic properties of the earth at the frequencies of the free oscillations. The distribution of the anelastic properties can be obtained by comparing the decay rate at various frequencies, since different frequencies are affected by different portions of the earth. Furthermore, several mechanisms of dissipation will be prominent in the various oscillations. The spheroidal oscillation of order 2 involves the entire earth, including the core, and the motion contains components of compression and shear. The radial oscillation ${}_0S_0$ involves only compressional motion; this oscillation provides a measure of the earth's dissipation in compression. The broadening of the toroidal lines is due primarily to dissipative processes within the mantle. The interaction of the core and mantle provides an additional sink of energy. But a detailed study of possible viscous and hydromagnetic effects rules out the core-mantle boundary as a major contributor to the energy loss.[42]

The half-width of the lines, Q, or the rate of energy dissipated per peak elastic energy, is found to be about 350 for spheroidal oscillations. Thus, in spheroidal oscillation the earth rings as a rather poor bell. The estimates of the Q for toroidal oscillations are not as good, somewhat lower figures being indicated, and the Q for ${}_0T_2$ is very uncertain. The highest Q of all is shown by the radial oscillation. As may be noted by comparing figures 5 and 6, the earth appears to be ringing in this mode of oscillation a month after the earthquake. The indicated Q is greater than 5000. Thus the dissipation due to compressional motion is much less than in shear.

The Problem of the Interaction of the Core and Mantle

The period of the fundamental toroidal oscillation is 42.9 minutes. This is more than a minute less than the period predicted for a theoretical model which gives a good fit to the other oscillation. The deviation is in the direction that one would expect if the core-mantle boundary were partly rigid; the resonant period for a mantle with a rigid inner surface is 32.1 minutes. A possible explanation for the apparent stiffness involves the earth's magnetic field. A component of the magnetic field tangential to the core-mantle boundary leaks out of the core into the conducting lower mantle. This component combines with the dipole component to give a Maxwell stress. The lower mantle is then partially glued to the core and this leads to an apparent stiffness. If this interpretation is correct, an estimate can be made of the conductivity of the lower mantle and the strength of the magnetic field. Thus detailed studies of the low frequency toroidal oscillations may lead to fundamental information regarding the electromagnetic properties of the core and mantle.

Section 5

THERMAL CONSTITUTION OF THE EARTH

The earth is a thermal engine. Any theory of the origin of the principal surface features of the earth must include assumptions or deductions for its thermal history. Since Kelvin's classic investigation, several important studies have been devoted to the determination of the earth's thermal character. Despite advances on many fronts, the distribution of temperature with depth remains uncertain, as does the principal mechanism by which heat is transported. At low temperatures, solids transport heat by ordinary thermal conduction. At high temperatures, large amounts of heat may be transported by radiation in optically transparent silicates. However, the combined effects of ordinary conduction plus radiation may be far exceeded by convection provided that the material making up the mantle can undergo mass motion. The problems of heat transport within the earth are thus closely tied to the question of the anelastic character of the

earth's material (see Section 2). If the earth behaves as an elastic solid with a certain critical stress, up to which only elastic deformations are possible, then thermal transport probably takes place solely by conduction and radiation. If, on the other hand, the material of the mantle, subjected to small stresses for a long time, undergoes creep, then convection is the dominant mechanism by which heat is transported. The ratio of heat transported by radiation and conduction to that transported by convection is $K/L\rho CV_r$, where K is the total conductivity including radiation, L is the scale length, ρ is the density, C is the heat capacity, and V_r is the radial velocity. Substituting numerical values we see that velocities on the order of 10^{-9} cm/sec are required for convection to dominate radiation and conduction. Because of the uncertainly in the mechanical behavior of the material, the nature of the thermal conditions in the earth remains most uncertain.

In the present chapter we consider the major factors which influence the internal distribution of temperatures. It is useful to base the discussion in terms of the problem of the development of the temperature distribution within the earth. Knowledge of the initial conditions are then required as well as the distribution of the heat sources in the earth and the radial variation of thermal conductivity, heat capacity, and density. The calculations then predict the present-day thermal distribution and surface heat flux.

A number of observations serve to limit the possible models of the earth's thermal history. The flow of heat from the interior of the earth averages about 50 ergs/cm^2-sec over both oceans and continents. Large regional variations exist but the impressive feature is the apparent equality of the average continental and oceanic heat flows. Seismology has established that the outer 2900 km of the earth is solid and that the material immediately underlying this mantle is fluid. A more accurate estimate of the melting points within the earth is now possible because of laboratory studies of the melting relations of geologically important materials at high pressures. The solid nature of the mantle thus limits the possible hypotheses of thermal distributions. In addition, the details of the variation of elastic velocity with depth and, in particular, the existence of a region of low velocity immediately under the crust, may provide a further indication of the thermal conditions within the earth.

TABLE 7.—*Half-Life and Heat Generation of Important Heat-Producing Isotopes*

Isotope	Half-life (years)	Heat generation (joules/gm-year)	Percentage of element's abundance
U^{238}	4.51×10^9	2.97	99.27
U^{235}	0.71×10^9	18.0	0.72
Th^{232}	13.9×10^9	0.82	100.0
K^{40}	1.25×10^9	0.92	0.0119

Age of the Earth

The age of the oldest rock provides a lower limit to the age of the earth. Gerling and Polkanov report a Pre-Cambrian rock evincing an age of 3.4×10^9 years.[43] The ratios of the stable lead isotopes have been used in numerous attempts to estimate the earth's age.[44] It has been established that the isotopic compositions of primordial lead in the earth appears to be identical to that in meteorites.[45] Patterson obtained the value of 4.5×10^9 years for the earth's age on the basis of lead isotopes.[46]

The age of chondritic meteorites has been determined by a number of workers by using both the strontium–rubidium method and the potassium–argon method. The ages scatter somewhat but are clustered around 4.0×10^9 to 4.4×10^9 years. For purposes of considering the earth's thermal history, the value of 4.5×10^9 years is here adopted as the age. This age is taken to be the time from the initial formation.

Heat Generation by Long-Life Radioactive Isotopes

Radioactive isotopes that have produced significant amounts of heat throughout the earth's history are distinguished by two characteristics: (1) The product of the abundance of the isotope and the rate of heat generation is relatively large; (2) The half-life of the isotope is of the same order as the age of the earth. The known isotopes that

TABLE 8.—*Heat Production and Surface Heat Flux for a Chondritic Earth*

Measurement	Chondritic Earth		Chondritic Mantle	
	Today	Initially	Today	Initially
Heat Production (ergs/sec)	3.01×10^{20}	24.3×10^{20}	2.04×10^{20}	16.5×10^{20}
"Equilibrium" Surface Heat Flow (ergs/cm²-sec)	59.0	476	40.2	324

meet these requirements are U^{238}, U^{235}, Th^{232}, and K^{40}. Table 7 lists the half-lives and heat generation of the important isotopes. Isotopes with shorter half-lives may have played an important role in the initial stages of the development of the earth.

Radioactive Composition of the Earth

In accordance with a suggestion by Urey, most discussions of the earth's thermal history assume that the earth and the chondrites have the same abundances of the radioactive elements. Strong supporting evidence for such an assumption follows a comparison of the present heat flow and production of heat that would exist in an earth of chondritic composition ($K = 8.0 \times 10^{-4}$ gm/gm, $U = 1.1 \times 10^{-8}$ gm/gm, $Th = 4.0 \times 10^{-8}$ gm/gm). If all the heat produced by radioactivity instantaneously reached the surface, then the present heat flow in a chondritic earth would be 59.0 ergs/cm²-sec (table 8). If only the mantle were chondritic then the present rate of heat production would be equivalent to a heat flux of 40.2 ergs/cm²-sec. The average of the observed heat flow is about 50 ergs/cm²-sec.

The relative production of heat, by the various isotopes, for a chondritic model of the earth is shown in table 9. At present, about 60 percent of the heat is due to K^{40}, U^{238}, and Th^{232}, each contributing about 20 percent. The relative heat production of K^{40} was much greater in the early stages of the earth's history.

TABLE 9.—*Heat Production by Isotopes*

Isotope	Today (ergs/gm-year)	4.5×10^9 Years Ago (ergs/gm-year)
U^{238}	0.32	0.64
U^{235}	0.014	1.14
Th^{232}	0.36	0.45
K^{40}	0.90	10.6

Surface Heat Flow Measurements

A most important quantity in any discussion of the thermal character of the earth is the amount of heat that is presently escaping from the earth's interior. The outward flux of heat by conduction per unit area per unit time is equal to the product of the thermal conductivity and the temperature gradient:

$$\frac{dQ}{dt} = -K\left(\frac{\partial T}{\partial r}\right). \qquad (32)$$

The observation of heat flow requires separate measurements of thermal conductivity and of the temperature gradient.

Systematic studies of thermal conductivities and temperature gradients have been carried out only since 1939. Birch reviewed the measurements on heat flow.[47] At that time (1954) about 40 determinations had been made on land areas and 25 at sea. The land areas showed a variation, roughly, of 20 to 120 ergs/cm²-sec. The variation in ocean measurements was somewhat greater, ranging from 15 to 140 ergs/cm²-sec. The mean of these measurements indicated an average heat flow of about 50 ergs/cm²-sec. Since that time the number of measurements at sea has more than doubled; and new measurements have also been made in continental areas. The new sea measurements indicate that the mean heat flow through the Pacific Ocean is about 50 ergs/cm²-sec[48] but

there is a far wider range in individual values, 10 to 370 ergs/cm²-sec. A partial explanation of this greater range in oceanic values may be that it is far easier to avoid thermally active regions on land than it is at sea. However, there is no simple explanation of the rather widespread lower values that have been found.

Thermal Conductivity of Silicate Materials

Birch and Clark demonstrated that the thermal conductivities of a wide variety of rocks show surprisingly small variation with composition.[49] Recent measurements at room temperature further establish the small variation.[50,51] Birch and Clark measured thermal conductivities up to 400°C and found that the conductivities of poor conductors such as feldspar and quartz increase with temperature, whereas the thermal conductivities of most materials decrease as the temperature increases. Studies of dielectric solids at somewhat higher temperatures show that thermal conductivity decreases approximately inversely with temperature, in accordance with phonon theory. No experimental studies have been carried out on the variation of thermal conductivity with pressure.

At high temperatures energy may be transferred within the solid by radiation rather than by lattice vibrations. Clark[52] shows that the contribution of radiation to thermal conductivity is approximately

$$K_R = 16\, n^2 \frac{sT^3}{3\epsilon} , \qquad (33)$$

where n is the effective index of refraction of the material, s is the Stefan–Boltzmann constant, T is the temperature in degrees Kelvin, and ϵ is the sum of the absorption and scattering coefficients averaged over all wavelengths. The striking feature of the contribution of radiation to conductivity is the strong temperature dependence. The higher the temperature the greater will be the amount of energy transferred by radiation, provided that the variation of the index of refraction and opacity do not overwhelm the T^3 dependence.

The opacity ϵ and the index of refraction are determined by the properties of the material, and in general they will be functions of temperature and pressure. The variation of the index of refraction of a given material with pressure has not been investigated experimentally. An estimate can be obtained by noting that, for materials of different compositions, the index varies linearly with density. If a law of corresponding states holds, then the index should vary linearly with pressure. Within the mantle the density varies from about 3.3 to 5.7, so that over the earth's mantle variation by a factor of 2 would be expected, with material at a depth having a higher index of refraction than that near the surface.

The rate at which radiation is transferred through a solid depends on the frequency of the radiation. Various mechanisms are known by which solids can absorb radiation, and these bebome important at different frequencies. Absorption due to excitation of lattice vibrations by radiation is strong in the infrared. This lattice absorption is relatively unimportant since at high temperatures the energy density is low at these long wavelengths. Intrinsic absorption is due to the excitation of valence electrons to the conduction band across the fundamental energy gap. Intrinsic absorption is important in the ultraviolet for wavelengths less than about 0.4 microns. The transparency of silicates to radiation is thus limited at long wavelengths by the infrared absorption due to lattice vibrations, and at short wavelengths by the absorption in the ultraviolet due to the excitation of electrons to the conduction band.

The region of high transparency in silicates lies in the visible to near infrared. This is not true for some of the conductors such as silicon and germanium which are opaque in the visible. Because of a lower energy gap the absorption edge in these materials lies in a much longer wavelength than in silicates. Within the visible and near infrared region of transparency, two processes can lead to absorption. Characteristic absorption peaks are associated with the presence of transition elements. The energy levels of the unfilled electron shells are split by the crystalline field and certain transitions between these split levels are allowed. These transitions lead to characteristic absorption bands of the transition elements. In silicates the most important transition element is iron, which has a strong absorption peak about one micron. Titanium, manganese, and other elements will

TABLE 10.—*Radiative Conductivities Calculated from Absorption Spectra*

Mineral	K_R (joules/cm-sec-deg)			
	1000°K	1500°K	2000°K	2500°K
Olivine, $(Mg, Fe)_2SiO_4$	0.297	0.86	1.45	2.02
Diopside, $CaMgSi_2O_6$	0.067	0.238	0.44	0.725
Pyrope, $Mg_3Al_2Si_3O_{12}$	0.004	0.021	0.075	0.184
Almandine, $Fe_3Al_2Si_3O_{12}$	0.004	0.017	0.042	0.212

also give rise to absorption bands. It is these bands that give color to the crystals.

The absorption between the peaks determines the contribution of radiation to thermal conductivity. If there is one perfectly transparent region, then the material has an infinite thermal conductivity. The general level of absorption between absorption bands limits the energy transported by radiation. This general absorption is primarily due to free electrons. In the classical theory, free electrons will absorb at all wavelengths. The dependence of the opacity on electrical conductivity is given by

$$60\pi\sigma/n, \tag{34}$$

where both the electrical conductivity σ and the index of refraction n vary with frequency. If we assume that the dc electrical conductivity is a sufficiently good approximation, the variation with opacity can be written as

$$\epsilon = \epsilon_0 + [60\pi\sigma_0 \exp(-E/kT)/n], \tag{35}$$

where ϵ_0 is the opacity at low temperatures and the second term on the right-hand side takes into account the temperature dependence of the conductivity; E is a characteristic excitation energy.

From the preceding discussion, it is apparent that the region of transparency in the near infrared and in the visible (between the ultraviolet absorption and the infrared lattice absorption) can be closed by: (1) the presence of absorption peaks due to transition elements; or (2) a high level of general absorption due to free carriers, and, in particular, free electrons. Furthermore, it must be expected that changes in temperature and pressure will affect the nature of the region of transmission.

A few direct measurements of thermal conductivities at high temperatures suggest the importance of radiation. Most estimates of the contribution of radiation are based on determination of the absorption spectra rather than on actual measurements. In a few cases the measurements of spectra can be made as a function of temperatures; these indicate the closing, at high temperatures, of the region of transparency for glasses containing transition elements. Clark measured the room temperature absorption spectra of several silicates in the visible and near infrared.[52] Radiative conductivities calculated from the absorption spectra are shown in table 10. It should be noted that the ordinary conductivities at room temperature are on the order of 0.02–0.03 joules/cm-sec-deg.

The absorption spectra of olivine and diopside are similar. Each has a peak at about one micron due to the ferric ion. The ultraviolet absorption edge lies at about 0.35 microns. Clark concluded that ferromagnesium silicates are sufficiently transparent for radiation to make an important contribution to the radioactivity. The results suggest a value of about 10 cm^{-1} or less as appropriate for the opacity of silicates at room temperature.

Data on the shift of absorption with an increase in temperature and pressure are still incomplete. It seems probable that increases both in temperature and pressure will tend to move the absorption edge toward longer wavelengths and reduce the region of transparency. An additional effect is the marked increase of electrical conductivity with an increase in temperature. If this increase is interpreted in terms of intrinsic conduction by free electrons, then the opacity will increase with

an increase in temperature. Equation 35 illustrates the exponential dependence of the opacity and temperature through the mechanism of electrical conductivity.

Melting Relations at High Pressure

The fact that the mantle is solid can be used to limit the possible temperature distribution if the dependence of the melting point on pressure is known. Data on the increase of the melting point with pressure are still scanty. Boyd and England[53] have determined the melting curves of diopside and albite ($NaAlSi_3O_8$). The measurements on diopside extend from 5 to 50 kilobars and those on albite from 10 to 32 kilobars. The melting curve of albite is intersected by the transition of albite to jadeite plus quartz at a pressure slightly in excess of 32 kilobars.

It is customary to represent the melting relations in terms of the Simon semiempirical equation. The pressure and temperature along a fusion curve are given by

$$P = \frac{a}{B}\left[\left(\frac{T}{T_0}\right)^B - 1\right] , \qquad (36)$$

where a and B are empirical constants. T_0 is the melting temperature at 1 bar pressure. The initial slope of the melting curve can be used to determine a, and B is determined by the initial curvature. Values for the coefficients in the Simon equation are given in table 11 for diopside and albite.

If the diopside melting curve is extrapolated to a pressure of 1.3×10^6 bars, the pressure of the core-mantle boundary, the melting point is 3900°C. It should be noted that the melting points give estimates of the possible maximum temperatures. The melting points refer to a system that can be treated as a single component. Naturally occurring silicate materials form multicomponent systems, and the melting points are lowered.

TABLE 11.—*Coefficients for the Simon Equation (see reference 53)*

Mineral	T_0 (° K)	Calculated Initial Slope (deg/kilobar)	a	B
Albite	1391	14	19.3	5.1
Diopside	1665	15	25.2	4.4

Calculation of Internal Thermal Conditions

The internal thermal state of the earth can be investigated by assuming an initial temperature distribution, a distribution of heat sources, and a distribution, with depth, of the parameters determining density, heat capacity, and thermal conductivity. The temperature at the surface of the earth will be assumed to be known and taken as constant in time. The problem is then to determine the distribution at later times. If volume changes associated with the changes in temperature are neglected, the equation governing the temperature distribution in a spherical symmetrical body is

$$\rho C_p \frac{\partial T}{\partial t}(r, t) = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 K \frac{\partial T}{\partial r}\right) + A(r, t) . \qquad (37)$$

The rate of heat production per unit volume, A, is a function of time as well as radius, since if a radioactive element is producing heat at a rate dQ/dt today, it produced heat at a rate $\exp[\lambda t(dQ/dt)]$ at a time t years ago; λ is the disintegration constant of the radioactive element. Analytical solutions to this equation can be obtained for special problems if the thermal conductivity is assumed to be independent of temperature.[54] If radiative transfer of heat is important, the conductivity depends on temperature and the equation becomes a nonlinear partial differential equation. In this case there are no known solutions for the general problem and numerical methods have been used.

Lubimova[55] and MacDonald[56] have carried out detailed calculations on the development of temperature within the earth. Several general features of the calculation are important. The principal effect of radiative transfer of energy on the distribution of temperature in the earth is the flattening of the temperature gradient with depth. Near the surface the temperature gradient is dependent only on the lattice conductivity of the solids, and a near-surface gradient on the order

TABLE 12.—*Present Surface Heat Flow in Differentiated Earth (reference 57)*

Heat source depth of burial (uniform distrib.) in km	Surface heat flow (ergs/cm²-sec)
100	69.4
200	79.4
300	83.2
400	85.7
500	87.8
600	89.7

of 10–20°C per km results in a wide variety of distributions of heat sources and thermal parameters. Because of the contribution from radiation the temperature required to remove heat from the deep interior decreases as the effect of conductivity increases with temperature. As a result the melting temperatures are most closely approached in the outer few hundred km of the earth. If the heat sources are distributed throughout the mantle the melting temperature is most closely approached or exceeded at depths on the order of 400–800 km. If the radioactivity is concentrated in the upper mantle (upper few hundred km) the melting point is approached or exceeded in the range 200–600 km.

The increase of opacity due to the increase of the number of free electrons at high temperatures does not influence the temperature distribution in a major way. Therefore the temperature distribution is relatively insensitive to the parameters that determine the electrical conductivity in the earth. The rapid increase of temperature in the outer few hundred km may contribute to a rapid rise of electrical conductivity in this region.

The prediction of today's heat flow for a number of earth models having a chondritic radioactive composition, and vastly different initial conditions and distributions of heat sources, falls within a factor of 2 of 50 ergs/cm²-sec. The contribution of initial heat to the present heat flow is sufficiently great that the heat flow in any model in which the earth has passed through a molten stage is greater than the observed heat flow. The concentration of heat sources near the surface further increases the heat flow. A differentiated earth in which the differentiation is the by-product of melting is inconsistent with the hypothesis of an earth with a total radioactivity equal to that of chondritic meteorites. This is illustrated in table 12 where the heat flow for various models of the earth is shown as a function of the depth to which the radioactive heat sources are concentrated. The parameters describing the models are shown in table 13.

A differentiated once-molten earth could contain a radioactivity of, at most, half that of the chondritic radioactivity and remain solid. This conclusion is independent of the mode of the transport of heat, unless substantial amounts of heat can be transported by convection at temperatures well below the melting point of the solids.

The coincidence of the present surface heat flow with the present rate of heat production in chondritic materials may be explained in a number of ways. Radioactivity corresponding to a chondritic earth is entirely concentrated within the upper few hundred km. The initial temperature of the earth must have been less than some 1000°C; and heat is transported both by convection and conduction-radiation. Alternatively, chondritic radioactivity is distributed so that an appreciable proportion of the heat sources lie well below 600 km. The average initial temperature must be low. The contribution of initial heat and the higher rate of heat production in the past compensate for the heat produced but not reaching the sur-

TABLE 13.—*Parameters for Models Listed in Table* 12

Opacity ϵ_0	10 cm^{-1}
Surface temperature	0°C
Index of refraction	1.7
Heat capacity	1.3 joules/gm-deg
Lattice conductivity	0.025 joules/cm-sec-deg

Depth (km)	Initial temperature (°C)
100	1880
200	1970
300	2050
500	2180
1000	2490
1500	2640
2000	2930

face. Still another possibility is that the radioactivity is, at most, half that of the chondrites. Heat sources can then be near the surface and the initial temperature relatively high.

A further feature of the thermal calculations is the shallow depth from which heat can be reaching the surface. Even with a high contribution of the radiation to the conductivity, heat produced at depths greater than about 1000 km does not reach the surface. The earth has a large thermal inertia and the time scale associated with the diffusion of heat within the earth is long compared with the time scale for the earth's history.

Section 6

INTERPRETATION OF THE VARIATIONS OF SEISMIC VELOCITIES IN THE MANTLE

Several features of the variation of seismic velocity with depth require explanation. First, there is the discontinuity between crustal material and material underlying the crust. This discontinuity can be interpreted either in terms of a relatively sharp change in the chemical composition of the material or in terms of a phase transition in which materials of low density transform to high density compounds.

Below the crust the velocity decreases with depth, reaching a minimum at a depth of 150 km. The decrease of velocity with depth can result either from the effect of changing temperature and pressure or from a change of chemical composition. Between 200 and 900 km the rate of the rise of velocity is much larger than below 900 km. The rapid rise of velocity again indicates a gradual change of composition, phase, or both. The discontinuity between the elastic mantle and the fluid core can also be interpreted in terms of phase transitions or a sharp break in chemical compositions. The present chapter discusses the laboratory evidence that bears on the problem of interpreting the variation of seismic velocities. It should be emphasized that the velocities shown in figure 1 are only average velocities and there is no reason to expect that material in the upper regions of the mantle is homogeneous. Indeed, the inhomogeneity of the near-surface material strongly suggests that the upper mantle, at least, will show major inhomogeneities.

The Crust-Mantle Boundary

Seismic investigations on crustal structure have shown that the average thickness of the normal continental crust is approximately 35 km. The compressional and shear wave velocities are about 6.2 and 3.6 km/sec, respectively, near the surface, and increase with depth to approximately 7.0 and 3.8 km/sec. The seismic velocities for the subcrustal mantle material center about 8.1 and 4.7 km/sec. The seismic results for the oceanic regions indicate that the discontinuities separating the crustal material from the mantle lie at depths approximately 10 to 12 km below the surface of the sea. In general, the oceanic crust is somewhat less than 5 km thick with seismic velocities for compressional waves of 6.3 to 7.0 km/sec. The rocks below the discontinuity have compressional wave velocities centering around 8.1 km/sec, although a larger variation in these velocities is observed in the oceanic mantle than in the continental mantle.

The interpretation of the seismic data in terms of chemical composition cannot be unique. The seismic data must be combined with other independent sources of data in order to obtain an interpretation of the composition of the crust and upper mantle. At present, the hypothesis of an oceanic crust of basaltic composition is compatible with both gravity and seismic results. Oceanic islands give no evidence for large amounts of any other material and hence it is generally assumed that the crust is predominantly basaltic. The continental crust is inferred to be of a far more complicated character. Surface outcrops indicate a wide variety of compositions and the observed variations of seismic velocities indicate a variation with depth. Subcrustal material in both oceanic and continental regions is generally assumed to be periodotite or similar rock. In this interpretation the crust-mantle boundary is the result of a sharp change in chemical composition, the crustal material having a greater proportion of Na, K, Ca, Al, and Si than the subcrustal material.

An alternative hypothesis about the nature of the seismic discontinuity between crustal and mantle material is that of phase transition. The phase transition would involve the breakdown of the feldspar to a denser form such as jadeitic pyroxene.

TABLE 14.—*Constants for Solid–Solid Transitions at High Pressures*

Constants	$NaAlSi_3O_8$ + $NaAlSiO_4$ (albite) (nepheline) → $2NaAlSi_2O_6$ (jadeite)	$NaAlSi_3O_8$ (albite) → $NaAlSi_2O_6$ + SiO_2 (jadeite) (quartz)	SiO_2 (quartz) → SiO_2 (coesite)	Al_2SiO_5 (sillimanite) → Al_2SiO_5 (kyanite)	Sapphirine + Enstatite (+Sillimanite?) → Pyrope
T = bp − c					
b (°C/bar)	0.054	0.049	0.089	0.11	0.11
c (°C)	54	273	1740	1020	780
ΔV (cm^3/gm)	−0.083	−0.075	−0.046	−0.035	–
ΔS (joules/gm-deg)	−0.17	−0.15	−0.051	−0.031	–

The principal rearrangement involved in such transitions is the change in the local coordination of aluminum. In phases stable at high temperature aluminum is found in four coordination (four nearest neighbors); in phases stable at high pressure, aluminum takes on six coordination. Associated with the structural rearrangement is a 10–15 percent change in density. The laboratory results shown in table 14 and figure 8 indicate that the low density albite and nepheline will transform to higher density jadeite at a pressure equivalent to a depth of 30 km in the crust, provided the temperature is on the order of 400°C. Other materials undergo transitions at nearly the same conditions.

The laboratory results are only suggestive since the exact thermal conditions at the crust-mantle boundary are not known, and since the experimental work has been carried out on chemically idealized systems. The effect of other components on the transition is still uncertain.

The simple suggestion that the seismic discontinuities represent a phase transition from a basalt-like material into a high pressure form faces serious difficulties. Since basalts form a multi-component system, a transition would be anticipated to occur over a wide range of temperatures and pressures rather than along a single curve on the temperature-pressure plane. Although the width of the transition zone is uncertain, it might be on the order of a few thousand bars; in this case the transition zone would be spread over several km or tens of km. The seismic evidence on the sharpness of the transition is still unclear, but most investigators would assume that the transition from crust to mantle takes place over a relatively sharp interval, a few km at most.

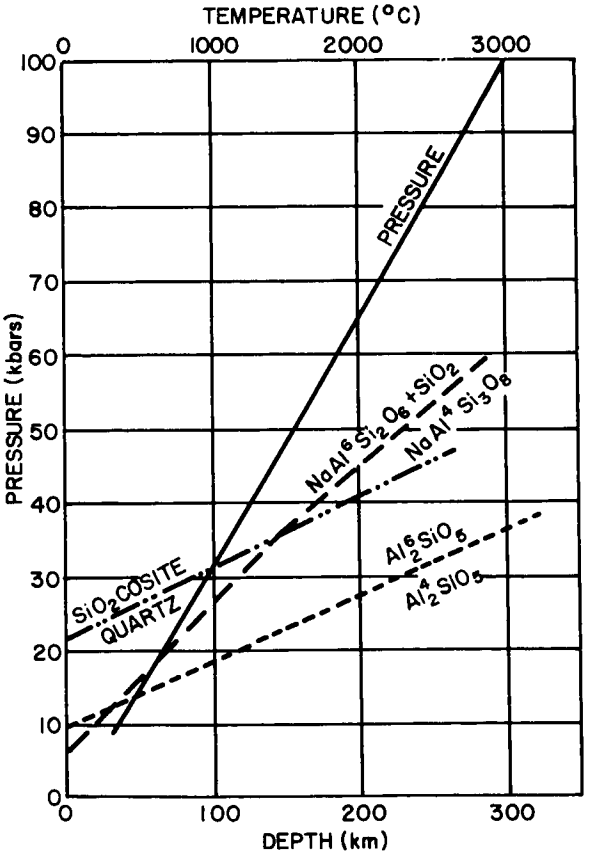


FIGURE 8.—Stability relations between sillimanite and kyanite (Al_2SiO_5), quartz and coesite (SiO_2), and albite ($NaAlSi_3O_8$), jadeite ($NaAlSi_2O_6$), and quartz.

An additional difficulty for a phase transition in oceanic areas follows from recent studies of the heat flow in the Pacific Ocean. Large regional variations in the surface heat flow correspond to variations in temperature of several hundred degrees at a depth of a few tens of km. Initial seismic studies over the same region show no comparable variation in the depth of the seismic discontinuity. Indeed the experimental data and assumed thermal gradients cannot be reconciled with the hypothesis of identical phase transitions under both oceans and continents. The depths of the seismic discontinuity can be explained in terms of a phase transition only if the crust and mantle have different compositions under oceans and continents.

Interpretation of the Low Velocity Layer

The physical conditions which give rise to the anomalous decrease in velocity below the crust are of great interest. Laboratory measurements show that in silicates the velocity increases with pressure; pressure stiffens a rock. An increase in temperature has the opposite effect, a decrease in the wave velocity. In the outer regions of the earth both pressure and temperature increase; thus velocity decreases if the increase in temperature is more effective than the increase in pressure. An extrapolation of laboratory data indicates that a gradient of 6° to 7°C per km is sufficient to produce a decrease in velocity.

If the low velocity zone is indeed due to the greater effect of the increase of temperature, then the existence of this zone indicates the magnitude of the temperature gradient in the upper regions of the mantle. The gradient between 30 and 150 km must be at least 6° to 7°C per km; and below 150 km the gradient must be less.

The low velocity zone might be due to large scale chemical inhomogeneity in the upper mantle. Presently available methods are insufficient to resolve the problem.

Structure of the Mantle between 200 and 900 Kilometers

Understanding the variation of velocity between 200 and 900 km is important in any discussion of the internal make-up of the inner planets. The terrestrial pressure at 200 km is 0.65×10^5 bars and this pressure is reached within Venus, Mars, and Mercury.

The present interpretation of seismic results proposes a gradual increase in velocity over the region of 200 to 900 km. Earlier, Jeffreys presented a solution with the discontinuity in velocity at the depth of 480 km.[58] Bernal proposed that the 480 km discontinuity was due to a phase change in which silicates transform from a lower density olivine structure to a higher density spinel structure. This suggestion was based on the similarity in chemical properties of silicon and germanium. The compound Mg_2GeO_4 was thought by Bernal to exist both in the olivine and in a higher density spinel structure. If silicon and germanium behave similarly, then olivine should also show a transition to spinel structure. The reality of the transition in Mg_2GeO_4 remained an open question until 1954 when Roy and Roy synthesized the high pressure, high density spinel phase.[59] In addition, recent results of high pressure experiments indicate other transitions.

Olivine and Spinel Structures

Olivine, $(Mg,Fe)_2SiO_4$, has a zero-pressure density ranging from 3.2 gm/cm³ for the pure magnesium end member to 4.3 gm/cm³ for the pure iron end member. The main features of the olivine structure are[60]:

1. The volume is largely determined by the arrangement of the oxygen. The oxygen atoms are in approximate hexagonal closest packing. They lie on sheets.
2. The silicon atoms are in tetrahedral coordination.
3. The magnesium or iron atoms are in octahedral coordination.

In spinel structure the oxygen atoms are approximately in cubic closest packing. In the interstices of the oxygen atoms are positions for metal atoms in four coordination between a tetrahedral group of oxygen atoms, and two sets of positions with an octahedral arrangement of nearest oxygen neighbors.

In the spinel structure one octahedron shares four edges with octahedra of the same kind and two with the tetrahedra. The other octahedron

shares one edge with a tetrahedron and two with other octahedra. The spinel thus differs from the olivine structure, in which each tetrahedron shares three edges with an octahedron and does not share the other three.

In the spinel form of Mg_2SiO_4 the magnesium atoms fill the octahedral positions and the silicon atoms fill the tetrahedral positions.[61] Each oxygen atom is linked to one silicon atom and to three magnesium atoms. The structure of the silicate spinel differs from the germanate, since in Mg_2GeO_4 the germanium occupies the six coordinated position and the magnesiums are split between the tetrahedral and octahedral positions.

Stability Relations for the Olivine-Spinel Transition

The stability relations between the olivine and spinel forms of Mg_2SiO_4 have been extensively studied by Dachille and Roy[61–63] and by Ringwood[64,65] They studied the temperature and pressure stability of solid solutions of $Mg_2(Si,Ge)O_4$. The transition of the pure iron end member has been similarly examined. Dachille and Roy found that the maximum silicate content of the spinel solid solution increases steadily with pressure from about 10 mole percent at 700 bars to 50 mole percent at 60,000 bars and 540°C.[61] They extrapolated these data to obtain the transition of the pure end member Mg_2SiO_4 at 100,000 bars, with an uncertainty of at least 15,000 bars. Ringwood synthesized the spinel form of Fe_2SiO_4 at 38,000 bars and 600°C.[66]

The volume change in the olivine-spinel transition is a matter of some uncertainty. It may be appreciably lower than the 8.2 percent volume change for the transition in Mg_2GeO_4. Dachille and Roy[61] estimate that the change in volume is 2.0 cm^3/mole or 4.6 percent, whereas Ringwood uses 4 cm^3/mole for the change in volume.

Dachille and Roy estimate that the slope of the transition curve is on the order of 13° per kilobar.[61] The presence of iron in the olivine will lower the transition pressure so that at temperatures on the order of 1000–2000°C an olivine with Mg/Fe = 9 would transform from the olivine form to the spinel form at depths on the order of 200–300 km.

Decomposition of Silicates to Oxides Under Pressure

A further possible transition involves the conversion of olivine or spinel forms of magnesium-iron silicate to oxide phases. Birch considered this possibility while noting that the ratio of the incompressibility to density (K_s/ρ) for a close-packed metal oxide such as MgO approaches the observed ratio of K_s/ρ within the earth at depths on the order of several hundred km.[67] At room pressure, the stable oxides chemically equivalent to Mg_2SiO_4 are quartz, SiO_2, and periclase, MgO. In periclase, the oxygen atoms are in approximate cubic closest packing with the magnesium and oxygen atoms at alternate corners of a set of cubes (NaCl structure). In quartz, each silicon atom is surrounded by four oxygen atoms. The tetrahedra of oxygen atoms, each with a central silicon atom, are joined together to form groups which combine in a low density, open framework.

The open framework of quartz is unstable at moderate pressures. At a temperature of 930°C, 30,000 atmospheres are required to convert quartz to coesite; coesite has a density of 3.02 gm/cm^3.[68,69] In coesite, each silicon atom has as its neighbors four oxygen atoms in tetrahedral arrangement, as in quartz.[70] The structure of coesite differs from that of quartz in having a more compact arrangement of the silica tetrahedra. Birch suggested still another possible form for SiO_2, the rutile structure assumed by TiO_2.[67] In this structure each silicon atom has as its nearest neighbors six oxygen atoms, approximately in the corners of a regular octahedron, and each oxygen atom is next to three silicon atoms, approximately at the corners of an equilateral triangle. The oxygen atoms are in approximate closest packing. The rutile structure thus assumes the general structure for AB_2 compounds, in which the A atom has six coordination. TiO_2 is also found in two other structures in which Ti is in six coordination but the oxygen atoms show slightly different arrangements. MacDonald, considering the transition of quartz to coesite, estimated the density of the rutile form of SiO_2 to be about 4.5–5.0 gm/cm^3.[68] Stishov and Popova synthesized the high pressure rutile form of SiO_2.[71] The density of the new compound was found to be 4.35 gm/cm^3. Chao, Fahey, et al.

later discovered the new high pressure form of SiO_2 in coesite-bearing rocks near Meteor Crater, Arizona, where the high pressure form, stishovite, formed during the impact of the meteor.[72]

The combination of SiO_2 and MgO is 21 percent denser than Mg_2SiO_4 in the olivine structure. The difference in density results from the fact that oxygen atoms in the two oxides are very nearly in closest packing with the largest change in density associated with the change in the coordination of silicon from four to six. The instability of the rutile form of SiO_2 at low temperature may be associated with the distance of nearest approach of the silicon atoms, each carrying a nominal charge of +4. In the quartz and coesite structures the silica tetrahedra are joined in such a way that each oxygen is shared by two silicon atoms. A pair of silicon atoms are shielded from each other by the presence of the intervening oxygen atom. In the rutile structure the octahedra surrounding the silicon atoms are joined together by shared edges. This results in the much closer approach of the silicon atoms without direct shielding of an oxygen atom. Silicon and titanium differ in behavior because titanium is a transition group element and the larger titanium ion forms bands that are less localized than those of silicon. Further, it may be noted that the anatase and brookite structures may be much less stable for SiO_2 than for TiO_2. In the anatase structure three edges are shared, and in both forms there is greater opportunity for the cations to interact directly with each other. Moreover, these other possible arrangements of octahedrally coordinated silicon lead to no saving in volume.

Stability Relations for the Oxide Transitions

Estimates of the possible conditions under which the breakdown of Mg_2SiO_4 to the oxides takes place depend upon recent, fragmentary data.

Stishov and Popova synthesized the rutile form of SiO_2 at a quoted pressure of 1.6×10^5 bars and a temperature between 1200 and 1400°C.[71] The estimate of pressure is probably high because friction in the apparatus was undervalued.[73] Let us assume that the coesite-stishovite transition pressure is 1.12×10^5 bars (30 percent less than the published value) and that the temperature is 1300°C. These values, and the thermochemical properties for the various phases listed in table 15, give a rough estimate of the equilibrium conditions for the reaction

$$Mg_2Si^4O_4 \rightarrow 2MgO + Si^6O_2, \qquad (38)$$

where the superscripts 4 and 6 on Si denote the coordination number. The change in internal energy in the transition of coesite to stishovite

TABLE 15.—*Thermochemical Properties for* MgO, SiO_2, *and* Mg_2SiO_4*

Material	Volume (cm^3/mole)	Density (gm/cm^3)	Entropy (joules/deg-mole)
MgO	11.2	3.59	27.2
Si^4O_2 (quartz)	22.6	2.65	41.8
Si^4O_2 (coesite)	19.9	3.02	38.9
Si^6O_2 (stishovite)	13.8†	4.35	28.0‡ - 33.0‖
Mg_2SiO_4 (olivine)	43.8	3.22	95.2
Mg_2SiO_4 (spinel)	39.8	3.37	–

*Reference 74.
†Reference 71.
‡Estimated on the basis of the "law of corresponding states" for quartz-coesite-stishovite.
‖Estimated on the basis of ≈ 2.4 joules/deg-cm^3 found for many close-packed oxides at temperatures above the Debye temperature.

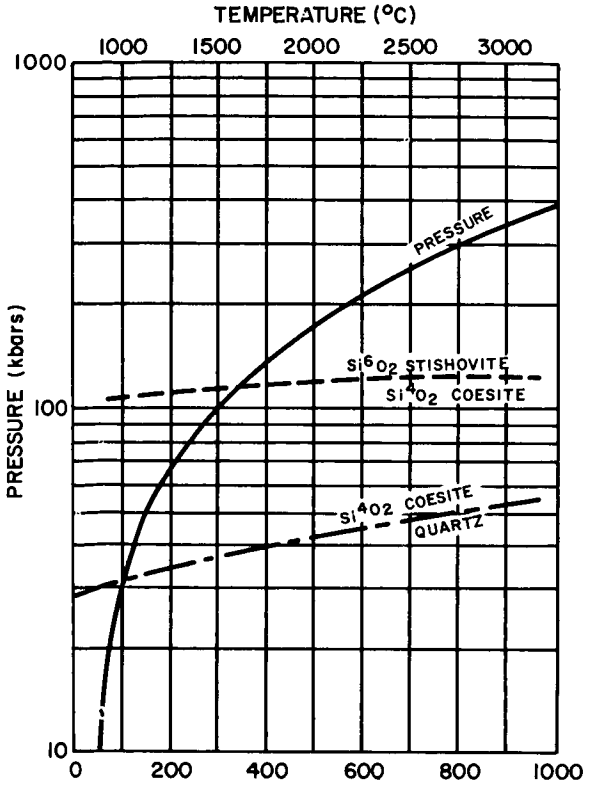


FIGURE 9.—Calculated stability relations between stishovite and coesite and stability relations between coesite and quartz extrapolated from data of Boyd and England (reference 69).

$(Si^4O_2 \rightarrow Si^6O_2)$ is

$$\Delta E = -P\Delta V + T\Delta S \approx (6.6 - 1.7) \times 10^4$$

$$= 4.9 \times 10^4 \frac{\text{joules}}{\text{mole}}, \qquad (39)$$

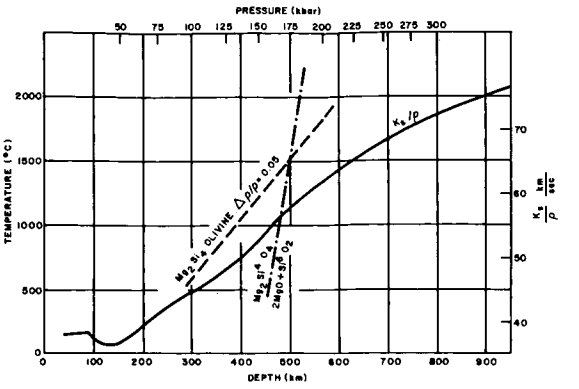


FIGURE 10.—Calculated curve for the breakdown of olivine to periclase plus stishovite. The variation of $\phi = K_s/\rho$ is derived from the Gutenberg model earth shown in figure 1.

for a heat of formation of Mg_2SiO_4, from the oxides, of 6.3×10^4 joules/mole; the change in internal energy for Reaction 38 is 11.2×10^4 joules/mole. The pressure required for the work term in equation 39 to equal $T\Delta S - \Delta E$ is then 1.7×10^5 bars at 1300°C. It should be noted that $T\Delta S$ for the transition is small compared to $P\Delta V$.

Estimated phase transition curves for the coesite-stishovite, olivine-spinel, and olivine-oxides transitions are shown in figures 9 and 10.

In the breakdown of the iron end member, Fe_2SiO_4, to the oxides, there is a volume change of 17.7 percent (table 16). Using thermochemical data listed for fayalite,[75] we find that the olivine to oxides transition pressure is about 10^5 bars. A summary of the volume changes associated with the various reactions is presented in table 16.

TABLE 16.—*Summary of Volume Changes for High Pressure Transitions in* Mg_2SiO_4 *and* Fe_2SiO_4

Material	Density of low pressure assemblage (gm/cm³)	Density of high pressure assemblage (gm/cm³)	Percentage change in volume
Mg_2SiO_4 (olivine)→Mg_2SiO_4 (spinel)	3.22	3.54*	11.0
Fe_2SiO_4 (olivine)→Fe_2SiO_4 (spinel)	4.3	4.85	12.0
$Mg_2Si^4O_4 \rightarrow 2MgO + Si^6O_2$	3.22	3.89	21.0
$Fe_2Si^4O_4 \rightarrow 2FeO + Si^6O_2$	4.3	5.06	17.7

*The values used are those given by Ringwood (reference 66). Dachille and Roy obtained a 4.6 percent volume change (reference 61). The origin of the discrepancy is not clear.

The temperature-pressure conditions under which high pressure transitions take place are still very uncertain. However, it is clear that under conditions existing at depths on the order of 200–500 km, the common silicate phases undergo transitions to new phases with structures in which the oxygen atoms lie in closest packing. Since the oxygen atoms occupy almost the entire volume within silicates, it is unlikely that the silicates will undergo further structural rearrangements. Any further decrease in volume can only result through a distortion of the outer electron orbits, i.e., pressure ionization. The pressure required for such transitions is on the order of 10^6 bars or greater, reached only within the cores of the larger planets. The density changes in these transitions are likely to be small compared to density shifts in the geometric rearrangements. The collapse of the silicate structures at a pressure of about 10^5 bars, however, can happen in Mars, Venus, and Mercury, since a pressure of this order is reached at moderate depths within the planets.

Chemical Composition of the Earth's Core

The discovery of the earth's core by Oldham in 1906 has provoked much speculation on its physical and chemical properties. The difficulty in determining the physical and chemical nature of the earth's core with any degree of certainty is connected with the difficulty of compressing materials to the pressure within the core. The pressure at the core-mantle boundary is 1.3×10^6 bars. This pressure is much greater than the highest pressure that can be obtained in static experiments at the earth's surface. However, it is possible to reach these pressures for short periods of time during the shock wave experiments, and data on the compressibility of solids obtained by using shock waves is of importance in establishing an equation of state for the core.

The core of the earth is generally assumed to be liquid iron–nickel alloy. Evidence advanced in support of this view includes: (1) the existence of iron meteorites; (2) the fact that the density of iron is in general agreement with the density of the core as estimated from seismology and from the theory of the figure of the earth; and (3) the fact that iron would probably be liquid at the temperatures and pressures of the core, whereas silicate materials would be solid. In addition, a recent investigation of the abundance of elements in the solar system suggests that iron is an abundant heavy element and a planet with an iron core would have the proper heavy element abundances.

The hypothesis of an iron core has not gone unchallenged. Kuhn and Rittmann propose a core of undifferentiated solar matter, arguing that an iron core is inconsistent with the relative abundances of the elements.[76] Ramsey suggests that the seismic discontinuity at 2900 km represents a change in phase from $(Mg,Fe)_2SiO_4$ in the olivine form in the mantle to the metallic phase in the core.[77–79] Birch critically reviews these suggestions and concludes that the core is principally iron–nickel with minor amounts of silicon and carbons, since the apparent density of the core is somewhat less than the expected density of pure iron.[67] The view that the 2900 km discontinuity is a phase change has many attractive features. However, a phase change in a multicomponent system will almost certainly be spread out over a range of pressures. The sharpness of the 2900 km discontinuity plus the large change in density at the core-mantle boundary favor a compositional discontinuity rather than a change of phase.

Estimates for the mean density of the core and for the density of the core at the core-mantle boundary follow from the estimates of the incompressibility, obtained from seismic velocities in the mantle (Section 4) and from the moment of inertia (Section 2). The satellite determination of J_2^0 leads to a much improved estimate of the mean moment of inertia, about 0.9 percent less than the value previously accepted. The new value of the moment of inertia implies a somewhat greater central concentration of mass. Table 17 lists the masses and moments of inertia of the principal structural units of the earth. The continental crust is taken to have a thickness of 35 km with a mean density of 2.8 gm/cm^3. The oceanic crust consists of a layer of 6 km of water underlain by 4 km of rock of a density of 3.0 gm/cm^3. The continent-ocean boundary is taken to be the 1000 fathom line.[19] The layer underneath the oceanic region extending to 35 km is assumed to have a density of 3.3 gm/cm^3.

TABLE 17.—*Masses and Moments of Inertia of the Major Structural Units of the Earth*

Structural Unit	Mass $\times 10^{-25}$ (gm)	Moment of Inertia $\times 10^{-43}$ (gm cm^2)
Crust	2.45	0.616
Continental	1.88	0.453
Oceanic	0.570	0.163
Layer Above 35 km	5.07	1.35
Mantle and Core	593	79.3
Mantle (below 35 km)	398	70.3
Core	194	8.99
Earth	597.7	80.68

The mass and moment of inertia of the mantle plus core follow by subtraction. Estimates of the individual moments of the core and mantle can be obtained from the Gutenberg velocity distribution (figure 1). The resulting mean density of the core is 11.28 gm/cm^3 and the density of the core-mantle boundary is 10.17 gm/cm^3.

Shock wave measurements on the equations of state of iron,[80,81] indicate that the density of iron at 1.3 megabars and about 2000°C is 11.2 gm/cm^3. Compare this value with the core-mantle boundary density of 10.17 gm/cm^3. The core material is about 10 percent less dense than iron. Nickel would raise the mean density whereas lighter elements, in particular silicon and carbon, would lower it. MacDonald and Knopoff suggest that perhaps silicon is the principal alloying element in the core.[82]

Section 7

INTERNAL CONSTITUTION OF THE MOON

Almost all inferences regarding the internal constitution of the moon follow from the data on its orbital and rotational motion. A possible additional limitation to the thermal constitution is derived from a comparison of the radio emission temperatures of the lunar surface at several wavelengths.[83]

The mean density is well established as 3.34 gm/cm^3. The density at 1 bar pressure and 25°C is uncertain because of the unknown thermal conditions and phase transformations within the moon. Urey estimates that the density is about 3.4 g/cm^3,[84] significantly less than the density of chondrites (3.57–3.76). He has pointed out that the moon's density is consistent with a material having about half the total iron content of the chondritic meteorites and has considered several chemical models. In particular, the abundance of the elements in the moon may be similar to the heavy element abundance in the sun's atmosphere. This possibility is illustrated in table 18 where the ratios of the chondritic to solar abundances are given. Note that iron is about 6 times, and potassium $\frac{1}{4}$, as abundant in chondrites as in the solar atmosphere.

The moon's density is almost equal to that of the upper part of the earth's mantle. The density

TABLE 18.—*Ratio of Chondritic to Solar Abundances for the Heavy Elements. Chondritic Abundances are the Averages of High and Low Iron Groups from Reference 85. Solar Abundances are from Reference 86.*

Element	Ratio	Element	Ratio
Na	0.77	Ca	1.23
Mg	1.59	Ti	1.52
Al	1.18	Cr	1.07
Si	1.00	Mn	2.24
P	0.66	Fe	6.03
S	0.17	Co	1.59
K	3.72	Ni	1.02

of the upper mantle is usually interpreted in terms of a periodotitic or eclogitic composition, the chemical composition of which is almost identical to the composition of the silicate phase of chondrites. The maximum pressure within the moon is about 46,000 bars; the pressure is too low for either the olivine-spinel or olivine-oxides transition.

The Moon's Gravitational Field

Urey emphasized that the figure of the moon sets limits on the internal constitution.[1] The basic argument is that the present gravitational figure of the moon is not in hydrostatic equilibrium. The inequalities in figure must be supported by internal strength. If the moon were partially molten or if a large part of it were nearly at the melting stage, then the figure should be close to hydrostatic. But the possible inferences about the internal thermal state are complicated by the fact that the stress differences may be supported in a sufficiently thick and cold outer shell, and the interior may be molten or nearly so. However, it is difficult to suggest a distribution of radioactivity that would lead to such a thermal configuration in a body the size of the moon. This point will be discussed in detail later.

We denote the moment of inertia about the axis of rotation by C. The least moment of inertia, A, is the moment about the axis pointing toward the earth. The moment B is along the tangent to the orbit. The ratios

$$\left.\begin{aligned} \alpha &= \frac{C-B}{A}, \\ \beta &= \frac{C-A}{B}, \\ \gamma &= \frac{B-A}{C}, \end{aligned}\right\} \qquad (40)$$

are determined from data on the moon's rotational and physical libration. The value for β depends on the angle of inclination between the mean axis of rotation and the pole of the orbit. Since the direction of the axis of rotation has been well observed, this ratio is accurately determined. A recent discussion by Jeffreys[87] gives

$$\beta = 0.0006270 \pm 0.0000015. \qquad (41)$$

The value of α depends on observations of the libration in longitude. Difficulties in observation have led to a wide scattering of results. Table 19 lists the relevant values.[87,88] Table 19 also lists the values for α and β, under the assumption that the moon is in hydrostatic equilibrium. The moon's equatorial bulge is much too large to be accounted for by its present rate of rotation and

TABLE 19.—*Data on the Moon's Figure* (*Reference* 87)

Computation	Observed Value	Theoretical Value		Ratio of Observed to Theoretical
		Hydrostatic Equilibrium	Uniform Density	
$\frac{C-B}{C}$	0.0004	0.0000094	–	42
$\frac{C-A}{C}$	0.0006279	0.0000375	–	16.7
$\frac{C-A}{Ma^2}$	0.000364	–	0.000251	1.45
$\frac{B-A}{Ma^2}$	0.000071	–	0.000084	0.84
$\frac{C}{Ma^2}$	0.56	–	0.4	1.4

the bulge toward and away from the earth is much greater than would be expected from tidal theory.

If the moon is supposed to be an elastic body supporting the inequalities in figure, the maximum stress differences at its center are on the order of 20 bars, provided the inequalities are supported by the whole body. Urey, Elsasser, and Rochester suggest that near-surface density inhomogeneities can lead to the observed figure.[89] The stress differences associated with these near-surface inhomogeneities will be larger than the 20 bars calculated for a homogeneous moon. It might be supposed that the outer layers of the moon are cold and could support the required stress differences. However, calculations of the possible temperature distribution within the moon bring the temperature close to the melting point at depths on the order of 300 to 400 km. The density inhomogeneities would then have to be concentrated in the outer 300 or 400 km.

Jeffreys proposed that the departure of the figure from equilibrium is due to the solidification of the moon at a time when it was much nearer to the earth.[20] The present figure then would represent a frozen tidal wave. However, on the tidal wave hypothesis the ratio α/β is independent of distance of the moon from the earth and has a theoretical value of 0.25. The observed ratio is 0.64.[87] Thus, the present figure cannot be explained solely by supposing that the moon solidified when it was much closer to the earth. A further difficulty follows from its thermal history. If the moon were initially molten, a uniformly distributed radioactivity only one-fourth that of chondrites would keep the moon near or at the melting point. The frozen tidal wave could not be maintained.

The ratios of the differences in moment of inertia to Ma^2 are determined by the orbital motion of the moon and in particular by the mean motion of perigee and node. The observed values are listed in table 19. Also listed are theoretical values obtained from α and β by assuming that the moon is of uniform density.[87] Combining the observed values of $(C-A)/Ma^2$ with β results in the ratio of the principal moment of inertia to the product of the mass and the square of the radius. For a uniform moon this ratio should be 0.4. The observed value is 0.56. Jeffreys considers the uncertainties in the observational material to be sufficiently large that the discrepancy with the theory for a homogeneous moon is not significant.

The data on rotation, libration, and orbital motion suggest a model which supports stress differences on the order of 10 bars, and in which the density distribution is either uniform or slightly increasing toward the surface. The presently available dynamical data are not consistent with a moon in which gravity-controlled density stratification on a large scale has taken place.

Lunar Tides

Both the earth and the sun produce significant tides on the moon. If the moon were spherical, the effect of these tides would be to distort the sphere into an ellipsoid whose long axis would be toward the tide-raising body. The moon rotates once per month relative to the line joining the centers of the sun and moon. Therefore, the long axis of the solar tidal ellipsoid rotates once per month relative to a system of axes fixed within the moon, giving rise to a predominantly fortnightly tide similar to the semidiurnal tides on the earth. The moon, however, does not rotate relative to the line joining the centers of the earth and the moon but wobbles about this line in a range of about 7 degrees. The tides due to the earth have, therefore, two causes: the changing distance of the earth which causes the ellipticity of the tidal ellipsoid to increase as the earth comes closer to the moon and decrease as the earth recedes, and the moon's wobble about the line of centers. The relative importance of these effects varies with the position of the moon's surface. Close to points of intersection of the line of centers with the moon's surface, the distance variation is the more important. The monthly variation of the earth-moon distance is the most important factor leading to large monthly tides.

The response of the moon to tide-raising potential can be calculated by using the Love numbers h and k. These Love numbers are determined by the distribution of density and rigidity in the moon's interior.

The moon's density of 3.34 gm/cm^3 is close to that of the upper parts of the earth's mantle and suggests that the moon is made out of silicate

materials similar to those in the mantle. A possible model of the moon, and the one usually considered, is a moon with a homogeneous composition. For a homogeneous incompressible spherical body the Love numbers are

$$\left.\begin{aligned} h &= \frac{5f}{2f+1}, \\ k &= \frac{3f}{2f+1}, \end{aligned}\right\} \qquad (42)$$

where $f=g\rho a/19\mu$, with g the surface gravity, a the outer radius, and μ the rigidity. If the density is taken as 3.34 gm/cm^3 and the rigidity as 7.38×10^{11} dynes/cm^2 (the rigidity of the upper mantle), we obtain $h=0.0331$ and $k=0.0199$, and the tidal gravity variation $(1+h-\frac{3}{2}k)$ is increased by 3.3 parts per thousand as compared with the variation on a perfectly rigid moon. If the moon were perfectly fluid, then h would equal 2.5 and k would be 1.5. If the moon has a fluid inner core with a core radius equal to half the outer radius then $h=0.0639$, $k=0.0384$, and there is a change in the moon's gravitational yielding of 6 parts in 1000.

The investigation of the elasticity of the moon by observations of the change of the tidal gravimetric acceleration is a possible experiment, but it is of marginal value because of the small effect.

Lunar Seismology

Seismic exploration of the moon will yield valuable information regarding its internal constitution. Prior to the actual landing of seismometers on the moon there are two important areas of investigation to consider. One is concerned with obtaining an estimate of eigen frequencies for the spheroidal oscillations. These long period oscillations can be detected by long period vertical seismometers, and they are an especially valuable source for information on the internal distribution of elasticity. A second investigation concerns elastic energy built up through differential heating and cooling associated with the internal release of radioactive heat and with the surface heat flow. The object in this case would be to estimate the magnitude of release of this elastic energy.

Computations of the expected spheroidal oscillations of the moon have been made by Bolt,[90] and by Takeuchi, Saito, and Kobayashi.[91] The results from reference 91 are listed in table 20 where the model of the moon is taken to be homogeneous, the elasticity being that of the upper mantle of the earth.

TABLE 20.—*Computed Periods of Spheroidal Oscillation of the Moon (Reference 91)*

Order	Period (min.)
2	14.7
3	9.9
4	7.7
5	6.4
6	5.5
7	4.8
8	4.3
9	3.9
10	3.6

MacDonald determined the average rate of increase of the principal stress differences for several thermal models of the moon.[92] This increase was on the order of 10 bars per million years. Using this figure we can estimate the average rate of release of strain energy, provided some assumption is made on the strength of the lunar materials. The rate of increase of strain energy will be a maximum when the stress differences are of the order of the strength. For a strength of 100 bars the maximum strain energy per unit volume is on the order of 1.4×10^3 ergs/cm^2. The rate of release of strain energy by failure at 100 bars maximum stress difference will be approximately 4×10^{24} ergs/year. At present, the moon should have a high degree of seismic activity, unless the radioactivity of the moon is concentrated in a thin layer near the surface or is far less than that of chondritic materials.

The degree of seismic activity varies among various models of the distribution of heat sources with depth. In the models considered in Reference 92 the greatest release of strain energy is at depths of 100 to 700 km. The assumption of deeply buried radioactivity implies deep foci for lunar seismic disturbances.

A determination of the average rate of seismic activity on the moon can be used to estimate

the amount of radioactivity on the moon and possibly its distribution.

Thermal Constitution of the Moon

Thermal Radiation from the Moon

Baldwin has made a study of the thermal radiation of the moon at radio wavelengths, in order to place a limit on the possible heat flow at the moon's surface.[83] High frequency radio and infrared emission determines the temperature at and within a few cm of the surface. Baldwin has carried out observations at 178 Mc (168 cm), where the depth of penetration is greater than for the shorter wavelength emission. He estimates that the depth of penetration is greater than 25 meters and may be on the order of 200 meters. The uncertainty is due to the sensitive dependence of the penetration depth on the unknown dielectric constant of the lunar surface material.

Baldwin obtains a temperature of 233 ± 8°K at 178 Mc. This temperature is within 25°K of various estimates of the surface temperature. The indicated temperature gradient, when combined with various estimates of thermal conductivity, gives upper limits to the heat flow ranging from 1 to 10 ergs/cm²-sec. Any limit in this range can be compared with the surface heat flow for a moon of chondritic composition. The present rate of heat production in a chondritic moon would be 3.7×10^{18} ergs/sec which is equivalent to an "equilibrium" surface heat flow of 9.6 ergs/cm²-sec. The extreme upper limit of the heat flow obtained from the study of the radio wave emission is then barely consistent with the assumption of a chondritic moon; on the whole a lower heat flow is indicated.

Calculations on the Thermal Structure of the Moon

The present thermal state of the moon is determined by the distribution of: (1) radioactive elements; (2) thermal conductivity; and (3) initial temperature. Of these parameters, the expected variations of the thermal conductivity exercise the least influence on the present temperature distribution; the distribution of radioactivity and of the initial temperature are much more important.

A number of calculations on the internal temperature distribution within the moon have been made with methods developed for the study of the thermal history of the earth.[1,56,93] If the moon were initially at a uniform temperature of 0°C, then the present temperature distribution (after 4.5×10^9 years) would be as shown in figure 11. The radioactivity is taken to be equal to that of chondritic meteorites and assumed to be uniformly distributed throughout the moon. For comparison, figure 11 also shows the melting point curves of diopside ($CaMgSi_2O_6$) and iron. The melting point of a multicomponent silicate material will be lower than that of diopside. The close approach of the diopside melting point curve and the calculated temperature indicates that much, if not all, of the moon would be near or at the melting point. A uniform chondritic composition results in temperatures above the estimated melting temperatures of silicates even if the assumed initial temperature is low.

A large number of other models have been constructed in which the radioactivity is uniformly distributed and different thermal parameters have been assumed.[88,92] The outstanding feature of the temperature-depth curves for models of the moon is the shallow depth at which the melting points are reached or exceeded. Since the moon is a relatively small body the central pressure is low, on the order of 46,000 bars. This low central pressure does not markedly raise the melting point of the possible constituents. The thermal conductivity is sufficiently low that only the outer portions of the moon are cooling to

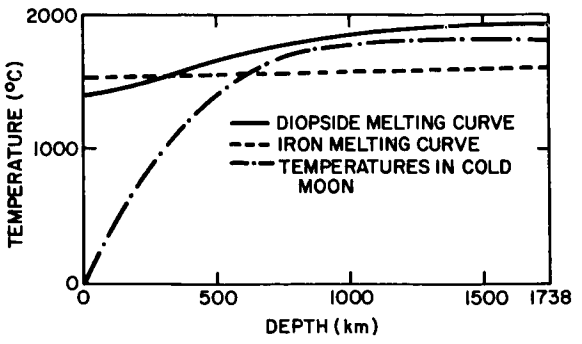


FIGURE 11.—Variation of temperature in a moon having chondritic radioactivity which is uniformly distributed. The initial temperature is 0°C.

any extent. The combination of these two circumstances results in the close approach of the melting point curve to the calculated temperature distribution illustrated in figure 11. In the earth the rate of increase of pressure with depth is much larger and raises the melting point, with the result that the earth's mantle is solid.

In a homogeneous moon, the melting temperature of silicates is approached or exceeded, provided that the moon has the composition of chondritic meteorites. The assumptions of chondritic radioactivity and of homogeneity are, then, inconsistent with the astronomical data on the figure of the moon, if these data are interpreted in terms of internal strength. There are two possible solutions: (1) The total radioactivity of the moon is less than that of chondritic meteorites. (2) The moon is a differentiated body whose radioactivity is concentrated toward the surface. In the latter case the suggestion of homogeneity in the orbital data must be disregarded. Both of these alternatives have been noticed by Urey;[93] he has carried out calculations on the effect of a lower radioactivity.

In investigating the two alternatives to a homogeneous chondritic moon, we will first consider the temperature distribution in a differentiated moon with the radioactivity concentrated toward the surface. Figure 12 shows the maximum

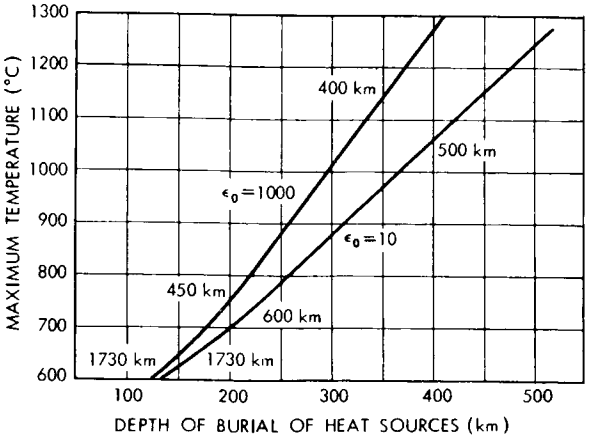


FIGURE 12.—Maximum temperature within the moon as a function of the depth of burial of heat sources. The depth at which the maximum temperature is reached is shown for the various depths of burial. The initial temperature is 600°C and the total radioactivity equals that of chondrites.

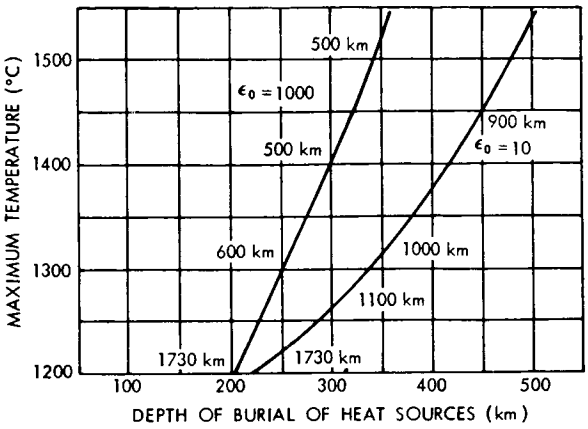


FIGURE 13.—Maximum temperature within the moon as a function of depth of burial of heat sources. This figure is similar to figure 12, but the initial temperature is 1200°C.

temperature and the depth at which this temperature is reached, for a moon in which the radioactivity is equal to that of chondritic meteorites but concentrated near surface layers. The initial temperature is taken to be 600°C. In the upper curve, the opacity is 1000 cm^{-1} and corresponds to a case where the radiative transport heat is unimportant. In the lower curve, the opacity is 10 cm^{-1} and radiation begins to dominate ordinary thermal conduction at temperatures on the order of 700°C. If the radioactivity is concentrated in the outer 100 km, then the maximum temperature is the initial 600°C since the central regions do not lose heat. These calculations illustrate that in a differentiated moon the melting point is not reached. The assumption of a differentiated moon removes the difficulty of a present-day partially molten moon, but raises the problem of how differentiation could proceed without the moon's passing through a molten stage. If the actual accretion of the moon is such as to permit a differentiated body, then the thermal difficulties are removed. However, there is no indication of density stratification in the orbital data.

In figure 13 the temperature is a function of the depth of burial of the heat sources, with an initial temperature of 1200°C. Even for this high initial temperature the maximum temperatures lie well below the melting temperatures at the depths involved.

Figure 14 shows the effect upon the maximum

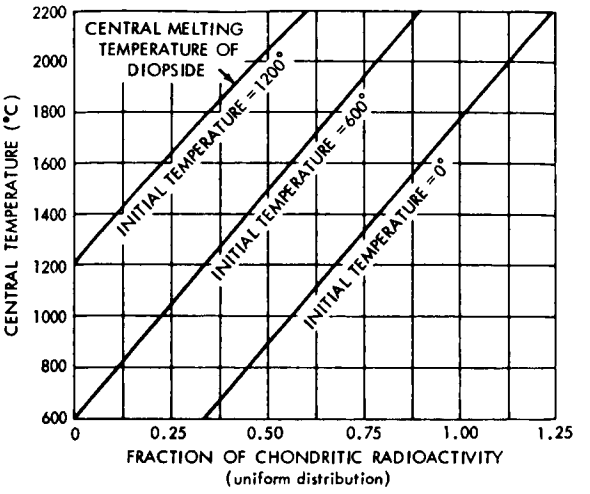


FIGURE 14.—The maximum temperature within the moon as a function of the total radioactivity of the moon. Radioactivity is uniformly distributed.

temperature of lowering the total radioactivity. The radioactivity is assumed to be uniformly distributed. In this case the maximum temperature is reached at the center of the moon. The central temperature is plotted as a function of the fraction of chondritic radioactivity maintaining the chondritic ratio among the heat-producing elements. It is to be noted that the solar abundance of potassium corresponds to about $\frac{1}{4}$ of the chondritic radioactivity. If the initial temperature is 0, the central temperature lies below the melting temperature for a total radioactivity equal to

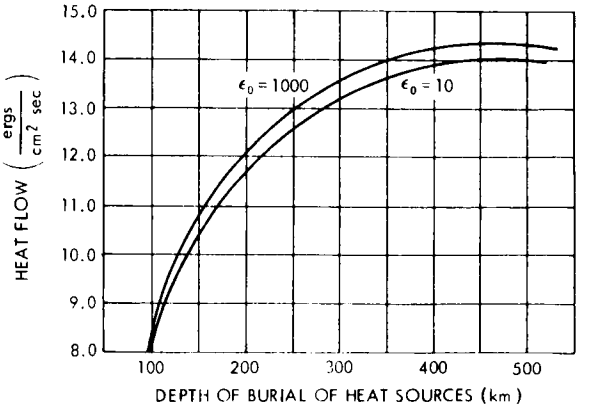


FIGURE 15.—The dependence of surface heat flow on the depth of burial of heat sources. Total radioactivity equals that of chondrites. The initial temperature is 600°C.

that of chondritic meteorites. If the initial temperature is 600°C, then the moon remains solid throughout if the radioactivity is 0.75 that of chondrites or less. An initial temperature of 1200°C requires that the radioactivity be less than 0.4 that of chondritic meteorites for the present temperature to lie below the melting point of diopside. If the radioactivity were as low as that indicated by the potassium content of the solar atmosphere, then the initial temperature could have been rather high, greater than 1200°C, and the lunar material would still have remained solid.

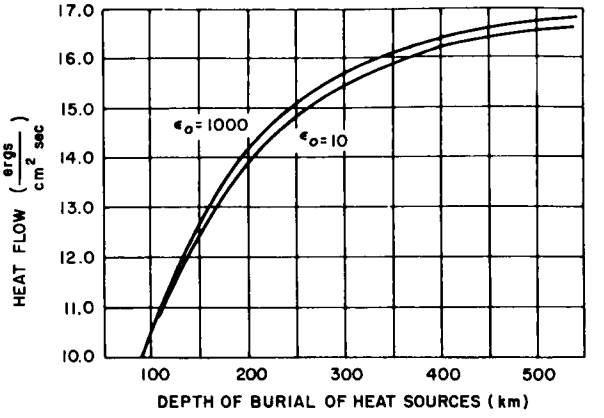


FIGURE 16.—Dependence of the surface heat flow on the depth of burial of heat sources. Total radioactivity equals that of chondrites. The initial temperature is 1200°C.

A differentiated moon removes the problem of the conflict between the inferred strength of the moon and the high temperatures within a uniform chondritic moon. Difficulties with the thermal emission may be raised in that the heat flow in a partially differentiated moon will be greater than in a uniform moon. Figure 15 shows the surface heat flow as a function of the depth of burial of heat sources, by assuming a chondritic radioactivity and an initial temperature of 600°C. Two values for the opacity are shown and these indicate that the detailed assumptions regarding thermal conductivity are not important in the determination of the surface heat flow. If the heat sources extend to a depth of some 500 km, then the surface heat flow is twice that of the model in which the heat sources are located within the upper 70 km. The near-surface concentration

of heat sources permits heat produced during the early stages of the history of the moon to escape rather than be trapped. If the heat sources are buried to a depth of some 500 km most of the heat is reaching the surface. This high heat flow may be in conflict with the results of radio emission at long wavelengths, but the point should not be held too strongly since the radio result involves rather arbitrary assumptions regarding the electrical properties of the near-surface lunar materials.

Figure 16 illustrates the dependence of the surface heat flow upon the depth of burial of sources when the initial temperature is at 1200°C. If the moon were initially molten, then the heat flow would be considerably higher, as in figure 17.

The alternative to concentration of heat sources near the surface is that the radioactivity of the moon is substantially less than that of chondritic meteorites. The dependence of surface heat flow on the amount of radioactivity is shown in figure 18 for two initial temperature distributions. In these models the radioactivity is assumed to be uniformly distributed throughout the moon.

The hypothesis that the moon is differentiated provides a solution to the discrepancy between its figure and the inferred high temperature of a moon of chondritic composition. There is no evidence in the astronomical data that the moon is indeed differentiated but the orbital data are not

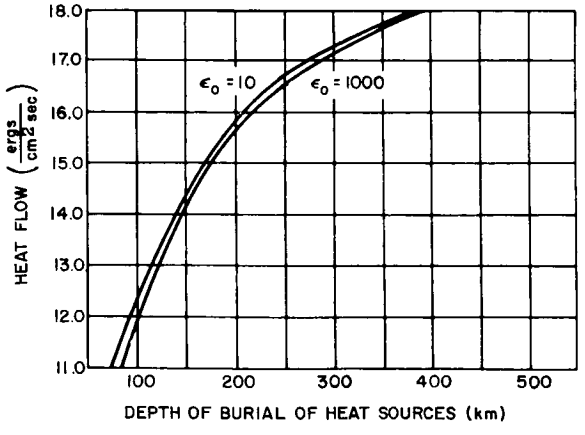


FIGURE 17.—Dependence of surface heat flow on the depth of burial of heat sources. Total radioactivity equals that of chondrites. Initial temperature is the temperature along the diopside melting curve shown in figure 11.

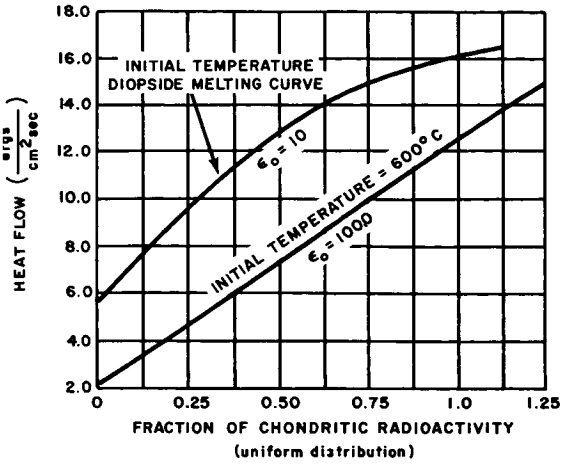


FIGURE 18.—Variation of surface heat flow with total radioactivity.

conclusive. The alternative supposition is that the radioactive element composition of the moon differs from that of meteorites. This hypothesis is strengthened by the fact that the density of the moon is appreciably less than the density of chondritic materials. The data on the moon can perhaps best be interpreted in terms of a heavy element composition like that of the solar atmosphere. This would explain the present density of the moon.[84] Further, the potassium content in the solar atmosphere is about a quarter that of the chondritic meteorites through this value is very uncertain. If the potassium content were indeed so low, and if the uranium and thorium were similarly reduced, then the radioactivity would be low enough to permit a chemically uniform moon without melting.

Section 8

INTERNAL STRUCTURE OF MARS

The mass of Mars is well determined from the orbital constants of the satellites Phobos and Deimos. Its density is uncertain because the diameter is disputed. The orbital constants for Phobos and Deimos also provide an accurate value of $(C-A)/Ma^2$. This ratio can be used to estimate the moment of inertia C, if hydrostatic equilibrium is assumed. One problem is then to construct such a model of the density distribution within the planet that the variation of density

TABLE 21.—*Radius and Mean Density of Mars*

Source	Radius (km)	Mean Density (gm/cm^3)
Rabe (Reference 94)	3415	3.84
Trumpler (Reference 95)	3310	4.21
Dollfus (Reference 96)	3335	4.12

yields the observed mass and moment of inertia. Since there is no definitive value for the radius, the problem is unsolved and there are a number of possible models of the density distribution within Mars.

The Gravitational Potential of Mars

Urey has discussed the difficult problem of obtaining the radius of the solid surface of Mars.[1] Rabe obtained a value of 3415 km;[94] this value is usually accepted as the correct visual radius. Later workers have shown that Mars exhibits a different radius for light at different wavelengths. Trumpler found that the radius in yellow light is 3310 km and obtained a value agreeing with that of Rabe for white light.[95] Dollfus arrived at a value 25 km greater.[96] Camichel, in a detailed study, obtained results which appear to confirm Trumpler's value.[97] Table 21 lists values for the radius and mean density of Mars obtained by the principal investigators. Though it appears that Trumpler's smaller radius value has been confirmed and is now generally accepted, we will show that it leads to difficulties in constructing a density-thermal model for Mars.

From Trumpler's value for the radius and Woolard's study of the motion of Phobos,[98] we obtain

$$J_2 = \frac{C - A}{Ma^2} = 0.00203 .$$

This value is obtained from observed quantities and does not depend on any assumptions regarding the attainment of hydrostatic equilibrium within the planet. By assuming the planet is in hydrostatic equilibrium it is possible to arrive at a value for the flattening, f (see equation 6), and, from the flattening, a value for C/Ma^2 (see equation 4). The hydrostatic flattening for Mars is 0.0052. It should be emphasized that this value is not the observed flattening, but rather the flattening calculated from J_2 with the assumption that the interior of Mars is in hydrostatic equilibrium.

In the earth the difference, ΔJ_2, between the observed value of 1.0823×10^{-3} and the hydrostatic value of 1.0711×10^{-3} requires that stress differences on the order of ten bars or greater are supported within the interior of the planet. The difference between the moon's radii as found from its dynamical effects is about 1.8 km and hardly any of it is explained by the hydrostatic hypothesis; it must be supported by the strength of the interior. The true value of the flattening of Mars may also differ from the hydrostatic value. In the earth the fractional difference between the observed and the theoretical flattening is

$$\frac{\Delta f}{f} = 0.0053 . \qquad (43)$$

Similar or even somewhat larger differences might be expected on Mars. On the earth the actual flattening is greater than the hydrostatic value and this has been interpreted to result from the secular deceleration of the earth (see Section 2 and references 19 and 24). For Mars the rate of rotational deceleration due to tidal interaction with Phobos is far less than for the earth. Nonetheless, we would expect the hydrostatic flattening on Mars to be less than the observed flattening. Percentage deviation can be larger on Mars since the maximum stress differences are proportional to gravity and gravity on Mars is about 0.38 that on the earth. On Mars $\Delta f/f$ might be on the order of 0.01 to 0.03 but probably is not much larger. The range of acceptable values of f is then about 0.005 to 0.0055. A much larger or much smaller value would lead to stress differences on the order of hundreds of bars rather than tens of bars and it is unlikely that any silicate material

can support these stress differences for long periods of time.

Density Distribution Within Mars

The construction of a density model for Mars requires the assumption of a surface density and a law of variation of density with pressure. To meet the requirements due to the total mass and J_2, an additional degree of freedom is necessary. Jeffreys suggested a chemically distinct core.[10] The free parameter to be determined for a given surface density is the core radius. In addition, the laboratory data and the interpretation of the structure of the earth's mantle suggest that phase transitions and major changes in density take place in silicates at pressures on the order of $1.0\times10^5-1.5\times10^5$ bars.

MacDonald has carried out detailed calculations on possible density distributions for Mars, assuming that the equation of state for the material is that for the material in the earth's mantle.[99] If a lower value for the radius of Mars is adopted, it would appear that the only model consistent with the dynamically obtained value of J_2 is one in which the surface density is large, about 3.8–4.0, and in which there is no large scale chemical inhomogeneity in the form of an inner metallic core. The indicated density is large compared with the density of the moon and chondrites. If the Trumpler value for the radius of Mars is adopted then the chemical composition of Mars must differ significantly from the chemical composition of the moon. An alternative possibility is that the outer radius is somewhat larger. In this case, the surface density of about 3.5 coupled with phase transitions would lead to an acceptable value of J_2. In such a model there could be a small core although its mass would be 1 percent or less that of the total body. The composition of Mars would then differ from the presumed chondritic composition of the earth in which the core contains 31 percent of the total mass of the earth. Possible density models are shown in figures 19 and 20. An additional restriction is introduced by the fact that Mars has an observable atmosphere. The atmosphere implies that the planet has undergone some sort of differentiation.

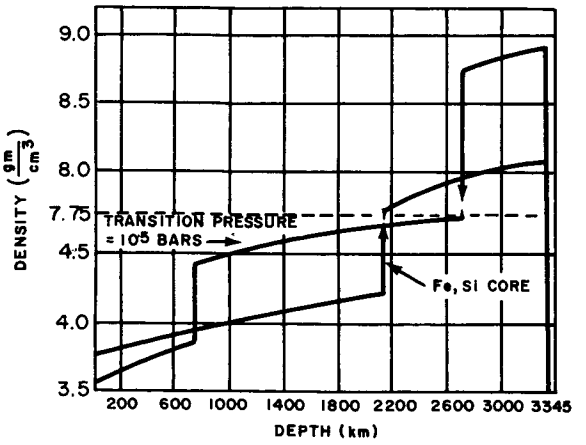


FIGURE 19.—Density models for Mars with an outer radius of 3345 km and a flattening of 5×10^{-3}. In one model phase transition takes place at a pressure of 10^5 bars and there is a relative change in volume of 15%. In the other model no phase transition is assumed. The mass of the chemically distinct core is 0.01 of the total mass in the phase transition model and 0.093 of the total mass in the other model.

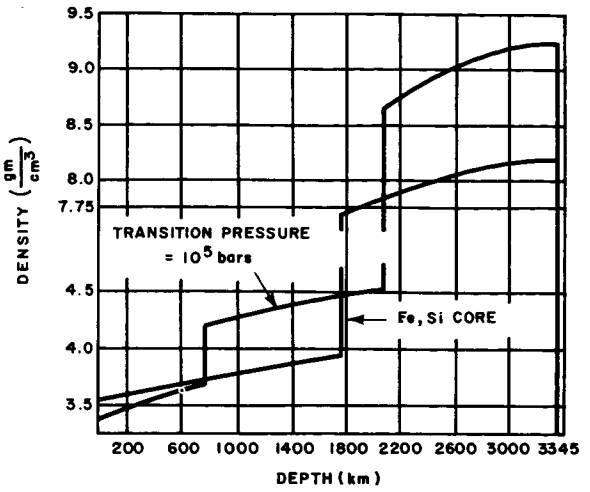


FIGURE 20.—Density models for Mars with an outer radius of 3345 km and a flattening of 4.70×10^{-3}. In one model phase transition takes place at a pressure of 10^5 bars and there is a relative change in volume of 15%. In the other model no phase transition is assumed. The mass of the core is 0.117 of the total mass in the phase transition model and 0.20 of the total mass in the other model.

Thermal Constitution of Mars

Detailed calculations have been carried out on the possible temperature distribution within Mars under the assumption that the composition

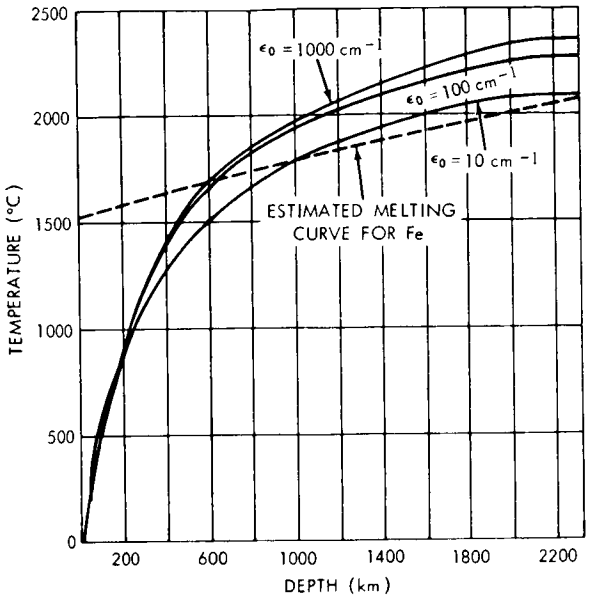


FIGURE 21.—Calculated temperature distribution in Mars I model (uniform chondritic radioactivity). The detailed parameters are given in table 22.

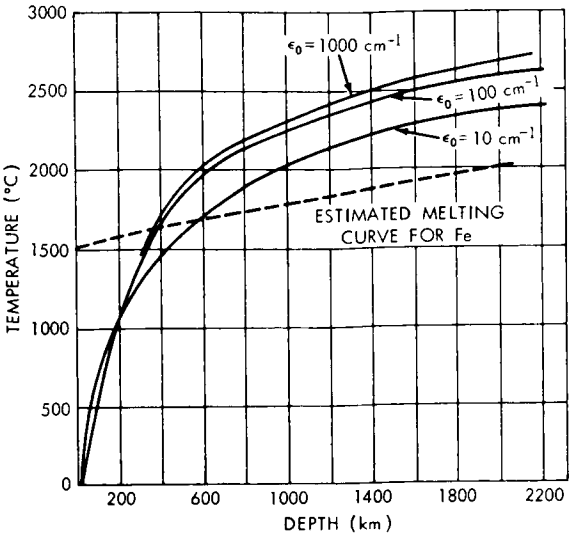


FIGURE 23.—Calculated thermal distribution in Mars III model (radioactivity concentrated in the outer two-thirds of the planet). Detailed parameters are given in table 22.

is chondritic. These calculations are illustrated in figures 21–24 for the parameters listed in tables 22 and 23. The outstanding feature of the temperature distribution is that if a chondritic radioactivity is assumed then the temperature exceeds the melting temperature of iron at shallow depths. Chondrites are about 12 percent metallic iron. If Mars were indeed of chondritic composition, the radioactivity would lead to the melting of the iron, and the iron would gravitationally differentiate to form a core composed of about 10 percent of the total mass of the planet. Such a

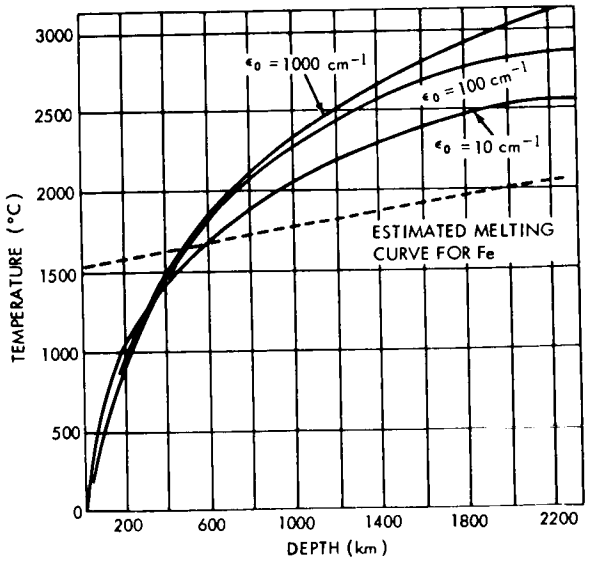


FIGURE 22.—Calculated temperature distribution in Mars II model (uniform chondritic radioactivity and an initial temperature twice as high as the model shown in figure 19). The parameters are shown in table 23.

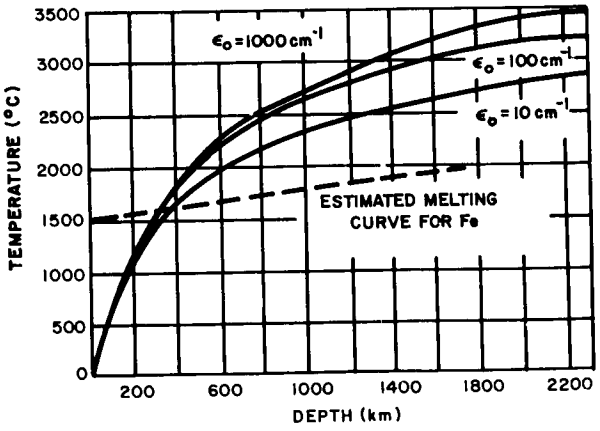


FIGURE 24.—Calculated thermal distribution in Mars IV, the model for the thermal parameters listed in table 23. The model is similar to that shown in figure 23 but the initial temperature is twice as high.

core is too large for the observed value of J_2, if the outer radius is 33.13 km.

The solution to the problem of the internal constitution of Mars awaits a definitive value for the radius of the solid body. From consideration of the mass and J_2 it would appear that the heavy element composition of Mars differs significantly from that of the moon and of the earth. The thermal calculations suggest a radioactivity substantially lower than that in chondrites.

TABLE 22.—*Thermal Parameters for Mars Models I and III (Figures 21 and 23, Respectively)*

Depth (km)	Initial temperature (°C)
0	−39
330	100
1040	400
1750	700
2000	800
2300	910
Radius (km)	3313
Surface temperature (constant in time), °C	−39

Radioactive element concentration (gm/gm)	Mars I	Mars III
Uranium	1.1×10^{-8}	1.35×10^{-8}
Thorium	4.4×10^{-8}	5.40×10^{-8}
Potassium	8.0×10^{-4}	9.8×10^{-4}

Time of integration	4.51×10^{9} yr.
Lattice conductivity	2.5×10^{5} (ergs/cm deg).
Density (uniform)	3.9 (gm/cm^3).
Heat capacity	1.46×10^{7} (ergs/gm deg).

Final surface heat flow (ergs/cm^2 sec)	Mars I	Mars III
$\epsilon_0 = 10$ cm	17.0	19.7
$\epsilon_0 = 100$ cm	12.5	16.0
$\epsilon_0 = 1000$ cm	12.0	15.0

TABLE 23.—*Thermal Parameters for Mars Models II and IV (Figures 22 and 24, Respectively)*

Depth (km)	Initial temperature (°C)
0	−39
330	200
1040	800
1750	1400
2000	1600
2300	1820
Radius (km)	3313
Surface temperature (constant in time), °C	−39

Radioactive element concentration (gm/gm)	Mars II	Mars IV
Uranium	1.1×10^{-8}	1.35×10^{-8}
Thorium	4.4×10^{-8}	5.40×10^{-8}
Potassium	8.0×10^{-4}	9.80×10^{-4}

Time of integration	4.51×10^{9} yr.
Lattice conductivity	2.5×10^{5} (ergs/cm sec deg).
Density (uniform)	3.9 (gm/cm^3).
Heat capacity	1.46×10^{7} (ergs/gm deg).

Final surface heat flow (ergs/cm^2 sec)	Mars II	Mars IV
$\epsilon_0 = 10$ cm^{-1}	19.3	23.5
$\epsilon_0 = 100$ cm^{-1}	14.2	17.6
$\epsilon_0 = 1000$ cm^{-1}	13.3	16.1

Section 9

ROTATION OF THE PLANETS AND THE THERMAL STRUCTURE OF VENUS AND MERCURY

The rates of rotation of Mercury and Venus indirectly provide information about the thermal conditions within these planets. The argument is based on the observations that the rates of rotation of six of the planets are remarkably similar despite major differences in their orbital characteristics, masses, and moments of inertia (see table 24). Mercury, Venus, and Pluto have ex-

TABLE 24.—*Rotation of the Planets*

Planet	Period (sec)	Mean angular velocity (radians/sec)
Mercury (reference 100)	7.6×10^{6}	8.3×10^{-7}
Venus	—	—
Earth	8.616409×10^{4}	7.292115×10^{-5}
Mars (reference 102)	8.86427×10^{4}	7.088×10^{-5}
Jupiter (reference 103)	3.54×10^{4}–3.57×10^{4}	1.77×10^{-4}
Saturn (reference 104)	3.68×10^{4}–3.83×10^{4}	1.67×10^{-4}
Uranus (reference 100)	3.85×10^{4}	1.63×10^{-4}
Neptune (reference 105)	4.58×10^{4}	1.37×10^{-4}
Pluto (reference 106)	5.5×10^{5}	1.14×10^{-5}

ceptionally low rates of rotation. Little is known of the possible dynamical history of Pluto and, since it may have been a satellite of Neptune at one time, it will not be considered further. Mercury's rate of rotation is uncertain but available evidence indicates a period equal to the orbital period.[100] The angular velocity of Venus is not known, but both spectroscopic and preliminary radar studies suggest that the rotational period is much longer than that of the earth.[101]

Rotation of Venus and Mercury

We will suppose that the *initial* angular velocity of rotation of Mercury and Venus equaled the *present* angular velocity of the major planets (a period in the order of 10 hours). On this hypothesis Mercury and Venus attained their present angular velocities through tidal interaction with the sun. Jeffreys has shown that the rate of change of the angular velocity of rotation should be proportional to the product of the sixth power of the planet's apparent diameter (seen from the sun) and $\sin\epsilon$, where ϵ is the phase lag of the solar tide.[20] The rate of change of angular velocity of Venus due to solar tidal interaction is about 7 times as great as that of the earth, provided the phase lag on Venus equals the phase lag on the earth. With equal phase lags, the present rate of deceleration of Venus would nearly equal that of the earth under the combined influences of the moon and the sun. The tidal deceleration of the earth due to the moon was greater in the past because of the closer approach of the moon. If the initial angular velocity of Venus were of the same order as that of the earth and if the internal properties of the two planets were similar, then the present angular velocity of Venus should be greater than that of the earth.

The apparent discrepancy can be resolved if it is assumed that the tidal dissipation takes place within the body of the planet and if anelastic dissipation in Venus is substantially greater than in the earth. In the earth, anelastic dissipation within the mantle is mainly due to solid friction.[24,80] At temperatures near or at the melting temperature other mechanisms (e.g., creep, elastoviscosity) may become dominant. If there were large regions within the interior of Venus at or near the melting point, then tidal dissipation would be greater than in the earth. A mixture of molten and solid silicates would be particularly effective in dissipation since the energy loss would be through direct viscous interaction.

The observed angular velocity of Venus may be explained by supposing that temperatures within the depths of the planet surpass the melting point over considerable regions. It appears likely that in the earth the temperature comes closest to the melting temperature at a depth on the order of 100 to 200 km, but below this the temperatures may well be below the fusion temperature. Pressure at a given depth in Venus is probably only a few percent less than the pressure at the same depth within the earth. The melting temperature of silicates at a given depth in Venus will be ap-

proximately equal to the corresponding melting temperature within the earth.

If the concentration of heat-producing radioactive elements in Venus is similar to that in the earth, the actual temperature at a given depth in Venus should be higher than in the earth. Extensive observations show that the microwave emission spectrum of Venus is flat between 3 and 21 cm, corresponding to a brightness temperature of 600°K.[107] If this high surface temperature has been maintained for an appreciable fraction of the history of the planet, then the internal thermal conditions have been affected in a major way. The higher surface temperature on Venus implies that the temperature in Venus would be higher than at the same depth in the earth.

A possible thermal structure within Venus is illustrated in figures 25 and 26. The basic data for these models are given in table 25. In Venus I the initial temperature (4.51×10^9 years ago) is taken as 1000°C, in Venus II it is 500°C. Three values of the opacity have been investigated. The range of ϵ_0 of 10 to 1000 cm^{-1} covers the values observed in silicates (see Section 5). The temperature at a depth of some 200 km is about 200°C higher than the temperature at the same depth for a similar thermal model of the earth. The variation of the melting temperature in silicates at these depths is uncertain, but the indicated temperatures are undoubtedly well above the melting temperature of common silicate materials.

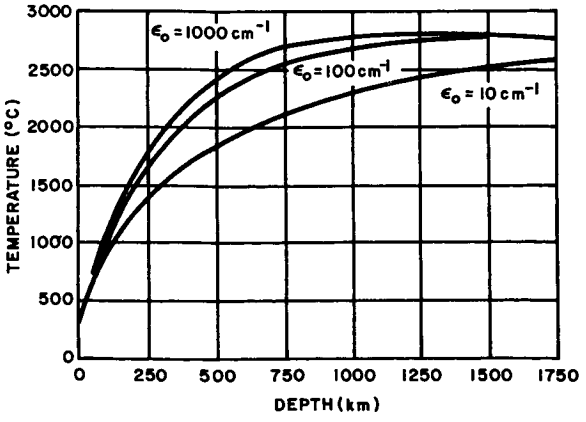


FIGURE 25.—Calculated temperature distribution in the Venus I model (initial temperature of 1000°C). Chondritic radioactivity is assumed to be uniformly distributed. The detailed thermal parameters are listed in table 25.

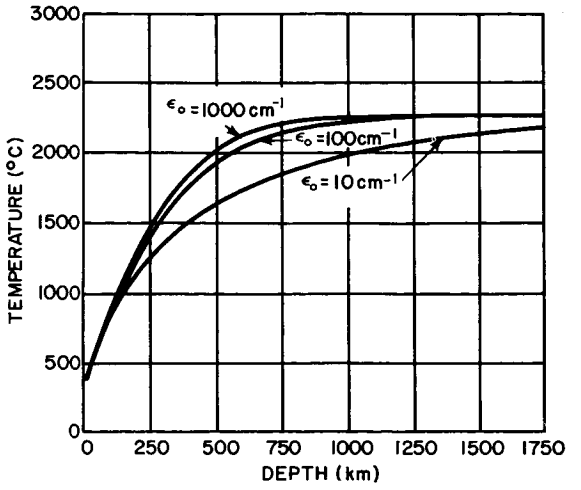


FIGURE 26.—Calculated temperature distribution in the Venus II model (initial temperature of 500°C). Chondritic radioactivity is assumed to be uniformly distributed. The detailed thermal parameters are listed in table 25.

The inferred low rate of rotation of Venus is compatible with a model in which a substantial fraction of the planet is at or near the melting point. The high surface temperature combined with a radioactive composition equal to that of chondritic meteorites would provide a consistent model. Although evidence is sketchy, present data are consistent with the hypothesis that the radioactive chemical composition of Venus and the earth are similar but that the high surface temperature of Venus has raised the temperature within the planet above the melting point of the silicates.

Inferences regarding the internal structure of Mercury are even more uncertain. The solar tide acting on Mercury is about 50 times more effective than the combined lunar and solar tides acting on the earth, provided that the phase lag is the same. If Mercury had an initial angular velocity similar to that of the *present-day* earth, then the time required to reduce its rate of rotation to the present value is much too long if its anelasticity is equal to that of the earth. On the hypothesis of an *initial* angular velocity *equal* to that of the earth, we would require either: (1) that Mercury was initially molten and during this early time

TABLE 25.—*Thermal Parameters for Venus Models I and II (Figures 25 and 26, Respectively)*

Initial temperature (uniform).	Venus I—1000°C; Venus II—500°C.
Surface temperature (constant in time).	300°C.
Radioactive element concentration (uniform, gm/gm):	
Uranium	1.1×10^{-8}
Thorium	4.4×10^{-8}
Potassium	8.0×10^{-4}
Time of integration	4.51×10^{9} yr.
Lattice conductivity	2.5×10^{5} (ergs/cm sec deg).
Density (uniform)	4.0 (gm/cm^3).
Heat capacity	1.3×10^{7} (ergs/gm deg).

Final surface heat flow (ergs/cm^2 sec)	Venus I	Venus II
$\epsilon_0 = 10$ cm^{-1}	17.2	14.4
$\epsilon_0 = 100$ cm^{-1}	17.7	14.6
$\epsilon_0 = 1000$ cm^{-1}	16.7	14.0

its periods of rotation and revolution were equalized by solar tidal friction; or (2) that the present thermal state of Mercury is such that it allows much greater dissipation of energy in it than in the earth. The thermal conditions within Mercury are most uncertain. The mean density makes it unlikely that the radioactive composition of Mercury is similar to that of chondrites. Indeed, it would seem that Mercury is largely metallic, and if this is the case then a lower concentration of radioactive elements is indicated. Since the effect of pressure on the melting point is far less in a planet the size of Mercury than in larger bodies such as Venus and the earth, a lower concentration of radioactive elements (one half to one third) would still melt the small body or make it largely molten.

REFERENCES

1. UREY, H. C., "The Planets; Their Origin and Development," New Haven: Yale University Press, 1952.
2. ALFVÉN, H., "On the Origin of the Solar System," Oxford: Clarendon Press, 1954.
3. LUST, R., and SCHLÜTER, A., "Drehimpulstransport durch Magnetfelder und die Abbremsung Rotierender Sterne," *Zeitschrift Astrophysik*, **38**(3): 190–211, December 12, 1955.
4. HOYLE, F., "On the Origin of the Solar Nebula," *Quart. J. Roy. Astronom. Soc.*, **1**(1): 28–55, September 1960.
5. CAMERON, A. G. W., "The Formation of the Sun and Planets," *Icarus*, **1**(1): 13–69, May 1962.
6. FOWLER, W. A., GREENSTEIN, J. L., and HOYLE, F., "Nucleosynthesis During the Early History of the Solar System," *Geophys. J.*, **6**(2): 148–220, February 1962.
7. REYNOLDS, J. H., "Determination of the Age of the Elements," *Phys. Rev. Letters*, **4**(1): 8–10, January 1, 1960.
8. ———, "Isotopic Composition of Primordial Xenon," *Phys. Rev. Letters*, **4**(7): 351–354, April 1, 1960.
9. ———, "I–Xe Dating of Meteorites," *J. Geophys. Res.*, **65**(11): 3843–3846, November 1960.
10. JEFFREYS, H., "The Density Distributions in the Inner Planets," *Monthly Not. Roy. Astronom. Soc., Geophys. Suppl.*, **4**(1): 62–71, January 1937.
11. O'KEEFE, J. A., ECKELS, A., and SQUIRES, R. K., "The Gravitational Field of the Earth," *Astronom. J.*, **64**(1272): 245–253, September 1959.
12. KOZAI, Y., "The Gravitational Field of the Earth Derived from Motions of Three Satellites," *Astronom. J.* **66**(1286):8–10, February 1961.
13. KING-HELE, D. G., "The Earth's Gravitational Potential, Deduced from the Orbits of Artificial Satellites," *Geophys. J.*, **4**, 3–16, 1961.
14. KAULA, W. M., "Celestial Geodesy," NASA Technical Note D-1155, March 1962; also in: *Advances in Geophysics*, ed. by H. E. Landsberg and J. Van Mieghem, New York: Academic Press, 1962, Vol. 9, pp. 191–293 (in press).
15. KING-HELE, D. G., "The Earth's Gravitational Potential, Deduced from the Orbits of Artificial Satellites," *Geophys. J.*, **6**(2): 270–272, February 1962.
16. NEWTON, R. R., "Ellipticity of the Equator Deduced from the Motion of Transit 4A," *J. Geophys. Res.*, **67**(1): 415–416, January 1962.
17. HENRIKSEN, S. K., "The Hydrostatic Flattening of the Earth," in: *Annals Internat. Geophys. Yr.*, New York: Pergamon Press, 1960, Vol. 12, Pt. 1, pp. 197–198.
18. O'KEEFE, J. A., "Zonal Harmonics of the Earth's Gravitational Field and the Basic Hypothesis of Geodesy," *J. Geophys. Res.*, **64**(12): 2389–2392, December 1959.
19. MUNK, W. H., and MACDONALD, G. J. F., "Continentality and the Gravitational Field of the Earth," *J. Geophys. Res.*, **65**(7): 2169–2172, July 1960.
20. JEFFREYS, H., "The Earth: Its Origin, History and Physical Constitution," 4th Ed., Cambridge: University Press, 1959.

21. Vening Meinesz, F. A., "Continental and Ocean-Floor Topography; Mantle-Convection Currents," *Proc. K. Nederlandse Akademie van Wetenschappen*, **63**B(4): 410–421, 1960.
22. Vacquier, V., Raff, A. D., and Warren, R. E. "Horizontal Displacements in the Floor of the Northeastern Pacific Ocean," *Bull. Geol. Soc. Amer.*,
23. Bullard, E. C., "The Figure of the Earth," *Monthly Not. Roy. Astronom. Soc., Geophys. Suppl.*, **5**(6): 186–192, January 1948.
24. Munk, W. H., and MacDonald, G. J. F., "The Rotation of the Earth," Cambridge: University Press, 1960.
25. Haskell, N. A., "The Motion of a Viscous Fluid Under a Surface Load," *Physics*, **6**(8): 265–269, August 1935.
26. ———, "The Motion of a Viscous Fluid Under a Surface Load. Part 2," *Physics*, **7**(2): 55–61, February 1936.
27. Lyustikh, E. N., "Isostasy and Isostatic Hypotheses," *Trudy Geologicheskogo Instituta, Akademiya Nauk SSSR No. 38* (165), 1957 (in Russian).
28. Chandler, S., "On the Variation of Latitude," *Astronom. J.*, **11**(251): 83–86, December 23, 1891.
29. Jeffreys, H., and Jeffreys, B., "Methods of Mathematical Physics," 3rd Ed., Cambridge: University Press, 1956.
30. ———, and Vicente, R. O., "The Theory of Nutation and the Variation of Latitude," *Monthly Not. Roy. Astronom. Soc.*, **117**(2): 142–161, 1957.
31. Bondi, H., and Gold, T., "On the Damping of the Free Nutation of the Earth," *Monthly Not. Roy. Astronom. Soc.*, **115**(1): 41–46, 1955.
32. Munk, W., and Hassan, E. S. M., "Atmospheric Excitation of the Earth's Wobble," *Geophys. J.*, **4**, 339–358, 1961.
33. Takeuchi, H., Saito, M., and Kobayashi, N., "Statical Deformations and Free Oscillations of a Model Earth," *J. Geophys. Res.*, **67**(3): 1141–1154, March 1962.
34. Bullen, K. E., "An Introduction to the Theory of Seismology," 2nd Ed., Cambridge: Univ. Press, 1953.
35. Gutenberg, B., "Physics of the Earth's Interior," New York: Academic Press, 1959.
36. Benioff, H., Press, F., and Smith, S., "Excitation of the Free Oscillations of the Earth by Earthquakes," *J. Geophys. Res.*, **66**(2): 605–619, February 1961.
37. Alsop, L. E., Sutton, G. H., and Ewing, M., "Free Oscillations of the Earth Observed on Strain and Pendulum Seismographs," *J. Geophys. Res.*, **66**(2): 631–641, February 1961.
38. Pekeris, C. L., Alterman, Z., and Jarosch, H., "Comparison of Theoretical with Observed Values of the Periods of Free Oscillation of the Earth," *Proc. Nat. Acad. Sci.*, **47**(1): 91–98, January 1961.
39. Ness, N. F., Harrison, J. C., and Slichter, L. B., "Observations of the Free Oscillations of the Earth," *J. Geophys. Res.*, **66**(2): 621–629, February 1961.
40. MacDonald, G. J. F., "The Earth's Free Oscillations," NASA Technical Note D–1336, June 1962; also *Science*, **134**(3491): 1663–1668, November 24, 1961.
41. Backus, G., and Gilbert, F., "The Rotational Splitting of the Free Oscillations of the Earth," *Proc. Nat. Acad. Sci.*, **47**(3): 362–371, March 1961.
42. MacDonald, G. J. F., and Ness, N. F., "A Study of the Free Oscillations of the Earth," NASA Technical Report R–136, 1962; also *J. Geophys. Res.*, **66**(6): 1865–1911, June 1961.
43. Gerling, E. K., and Polkanov, A. A., "The Absolute Age Determination of the Precambrian of the Baltic Shield," *Geokhimiya* No. 8: 695–717, 1958; Translation in *Geochemistry* No. 8: 867–896, 1958.
44. Russell, R. D., and Farquhar, R. M., "Lead Isotopes in Geology," New York: Interscience, 1960.
45. Murthy, V. R., and Patterson, C. C., "Primary Isochron of Zero Age for Meteorites and the Earth," *J. Geophys. Res.*, **67**(3): 1161–1167, March 1962.
46. Patterson, C., "Age of Meteorites and the Earth," *Geochim. et Cosmochim. Acta*, **10**(4): 230–237, October 1956.
47. Birch, F., "The Present State of Geothermal Investigations," *Geophysics*, **19**(4): 645–659, October 1954.
48. Von Herzen, R., "Heat-Flow Values from the South-Eastern Pacific," *Nature*, **183**(4665): 882–883, March 28, 1959.
49. Birch, F., and Clark, H., "The Thermal Conductivity of Rocks and its Dependence Upon Temperature and Composition," *Amer. J. Sci.*, **238**(8): 529–558, August 1940; **238**(9): 613–635, September 1940.
50. Clark, S. P., Jr., and Niblett, E. R., "Terrestrial Heat Flow in the Swiss Alps," *Monthly Not. Roy. Astronom. Soc., Geophys. Suppl.*, **7**(4): 176–195, August 1956.
51. Birch, F., "Flow of Heat in the Front Range, Colorado," *Bull. Geol. Soc. Amer.*, **61**(6): 567–630, June 1950.
52. Clark, S. P., Jr., "Absorption Spectra of Some Silicates in the Visible and Near Infrared," *Amer. Mineralogist*, **42**(11–12): 732–742, November-December 1957.
53. Boyd, F. R., and England, J. L., "Effect of Pressure on the Melting of Diopside and Albite," *J. Geophys. Res.*, **67**(9): 3544, August 1962 (Abstract).
54. Carslaw, H. S., and Jaeger, J. C., "Conduction of Heat in Solids," 2nd Ed., Oxford: Clarendon Press, 1959.
55. Lubimova, H. A., "Thermal History of the Earth With Consideration of the Variable Thermal Conductivity of its Mantle," *Geophys. J.*, **1**(2): 115–134, June 1958.

56. MacDonald, G. J. F., "Calculations on the Thermal History of the Earth," *J. Geophys. Res.*, **64**(11): 1967–2000, November 1959.
57. ———, "Surface Heat Flow from a Differentiated Earth," *J. Geophys. Res.*, **66**(8): 2489–2493, August 1961.
58. Jeffreys, H., "The Structure of the Earth Down to the 20° Discontinuity," *Monthly Not. Roy. Astronom. Soc., Geophys. Suppl.*, **3**(9): 401–422, April 1936.
59. Roy, D. M., and Roy, R., "An Experimental Study of the Formation and Properties of Synthetic Serpentines and Related Layer Silicate Minerals," *Amer. Mineralogist*, **39**(11): 957–975, November–December 1954.
60. Bragg, W. L., "Atomic Structure of Minerals," Ithaca, N.Y.: Cornell University Press, 1937.
61. Dachille, F., and Roy, R., "High Pressure Studies of the System Mg_2GeO_4–Mg_2SiO_4 with Special Reference to the Olivine-Spinel Transition," *Amer. J. Sci.*, **258**(4): 225–246, April 1960.
62. ———, and ———, "System Mg_2SiO_4–Mg_2GeO_4 at 10,000, 60,000 and about 300,000 psi," *Bull. Geol. Soc. Amer.*, **67**(12, pt. 2): 1682–1683, December 1956 (Abstract).
63. ———, and ———, "Experimental Study of the Olivine-Spinel Inversion in Mg_2SiO_4," *Bull. Geol. Soc. Amer.*, **69**(12, pt. 2): 1550, December 1958 (Abstract).
64. Ringwood, A. E., "The System Mg_2SiO_4–Mg_2GeO_4," *Amer. J. Sci.*, **254**(11): 707–711, November 1956.
65. ———, "The Constitution of the Mantle–II. Further Data on the Olivine-Spinel Transition," *Geochim. et Cosmochim. Acta*, **15**(1/2): 18–29, 1958.
66. ———, "Olivine-Spinel Transition in Fayalite," *Bull. Geol. Soc. Amer.*, **69**(1): 129, January 1958.
67. Birch, F., "Elasticity and Constitution of the Earth's Interior," *J. Geophys. Res.*, **57**(2): 227–286, June 1952.
68. MacDonald, G. J. F., "Quartz-Coesite Stability Relations at High Temperatures and Pressures," *Amer. J. Sci.*, **254**(12): 713–721, December 1956.
69. Boyd, F. R., and England, J. L., "The Quartz-Coesite Transition," *J. Geophys. Res.*, **65**(2): 749–756, February 1960.
70. Zoltai, T., and Buerger, M. S., "The Crystal Structure of Coesite, the Dense, High-Pressure Form of Silica," *Zeitschrift für Kristallographie*, **111**(2): 129–141, February 1959.
71. Stishov, S. M., and Popova, S. V., "A New Dense Modification of Silica," *Geokhimiya* No. 10: 837–839, 1961; Translation in *Geochemistry* No. 10: 923–926, 1961.
72. Chao, E. C. T., and Fahey, J. J., et al., "Stishovite, SiO_2, a Very High Pressure New Mineral from Meteor Crater, Arizona," *J. Geophys. Res.*, **67**(1): 419–421, January 1962.
73. Kennedy, G. C., and LaMoei, P. N., "Some Fixed Points on the High Pressure Scale," in: *Progress in Very High Pressure Research: Proc. Internat. Conf., Bolton Landing, New York, June 1960*, ed. by F. P. Bundy, W. R. Hibbard, Jr., and H. M. Strong, New York: Wiley, 1961, pp. 304–313.
74. MacDonald, G. J. F., "A Critical Review of Geologically Important Thermochemical Data," Doctoral Dissertation, Harvard University, 1954.
75. King, E. G., "Heats of Formation of Manganous Metasilicate (Rhodonite) and Ferrous Orthosilicate (Fayalite)," *J. Amer. Chem. Soc.*, **74**(17): 4446–4448, September 5, 1952.
76. Kuhn, W., and Rittmann, A., "Über den Zustand des Erdinnern und seine Entstehung aus einem Homogenen Urzustand," *Geologische Rundschau*, **32**(3): 215–256, August 25, 1941.
77. Ramsey, W. H., "On the Constitution of the Terrestrial Planets," *Monthly Not. Roy. Astronom. Soc.*, **108**(5): 406–413, 1948.
78. ———, "On the Nature of the Earth's Core," *Monthly Not. Roy. Astronom. Soc., Geophys. Suppl.*, **5**(9): 409–426, October 1949.
79. ———, "The Planets and the White Drawfs," *Monthly Not. Roy. Astronom. Soc.*, **110**(5): 444–454, 1950.
80. Knopoff, L., and MacDonald, G. J. F., "An Equation of State for the Core of the Earth," *Geophys. J.*, **3**(1): 68–77, March 1960.
81. Birch, F., "Composition of the Earth's Mantle," *Geophys. J.*, **4**, 295–311, 1961.
82. MacDonald, G. J. F., and Knopoff, L., "On the Chemical Composition of the Outer Core," *Geophys. J.*, **1**(4): 284–297, December 1958.
83. Baldwin, J. E., "Thermal Radiation From the Moon and the Heat Flow Through the Lunar Surface," *Monthly Not. Roy. Astronom. Soc.*, **122**(6): 513–522, 1961.
84. Urey, H. C., "Lines of Evidence in Regard to the Composition of the Moon," in: *Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January 1960*, ed. by H. K. Bijl, Amsterdam: North-Holland Publ. Co., 1960, pp. 1114–1122.
85. ———, and Craig, H., "The Composition of the Stone Meteorites and the Origin of the Meteorites," *Geochim. et Cosmochim. Acta*, **4**(1/2): 36–82, 1953.
86. Aller, L. H., "Solar and Stellar Abundances of the Elements," in: *Physics and Chemistry of the Earth*, ed. by L. H. Ahrens, F. Press, et al., London: Pergamon Press, 1961, Vol. 4, pp. 1–26.
87. Jeffreys, H., "On the Figure of the Moon," *Monthly Not. Roy. Astronom. Soc.*, **122**(5): 421–432, 1961.
88. MacDonald, G. J. F., "Interior of the Moon," *Science*, **133**(3458): 1045–1050, April 7, 1961.
89. Urey, H. C., Elsasser, W. M., and Rochester, M. G., "Note on the Internal Structure of the Moon," *Astrophys. J.*, **129**(3): 842–848, May 1959.
90. Bolt, B. A., "Spheroidal Oscillations of the Moon," *Nature*, **188**(4757): 1176–1177, December 31, 1960.

91. TAKEUCHI, H., SAITO, M., and KOBAYASHI, N., "Free Oscillations of the Moon," *J. Geophys. Res.*, **66**(11): 3895–3897, November 1961.
92. MACDONALD, G. J. F., "Stress History of the Moon," *Planet. Space Sci.*, **2**(4): 249–255, August 1960.
93. UREY, H. C., "Boundary Conditions for Theories of the Origin of the Solar System," in: *Physics and Chemistry of the Earth*, ed. by L. H. Ahrens, F. Press, et al., London: Pergamon Press, 1957, Vol. 2, pp. 46–76.
94. RABE, W., "Untersuchungen über die Durchmesser der Grossen Planeten," *Astronomische Nachrichten*, **234**(5606): 153–200, 1928.
95. TRUMPLER, R. J., "Observations of Mars at the Opposition of 1924," *Lick Observ. Bull.*, **13**(387): 19–45, April 26, 1927.
96. DOLLFUS, A., "Un Micromètre à Double Image Permettant un Grand Dedoublement," *Comptes Rendus, Academie des Sciences* (Paris), **235**(23): 1477–1480, December 10, 1952.
97. CAMICHEL, H., "Détermination Photographique du Pôle de Mars, de son Diamètre et des Coordonnées Aréographiques," *Bulletin Astronomique*, **18**(2): 83–174, 1954.
98. WOOLARD, E. W., "The Secular Perturbations of the Satellites of Mars," *Astronom. J.*, **51**(1150): 33–36, August 1944.
99. MACDONALD, G. J. F., "On the Internal Constitution of the Inner Planets," *J. Geophys. Res.*, **67**(7): 2945–2974, July 1962.
100. RUSSELL, H. N., DUGAN, R. S., and STEWART, J. Q., "Astronomy; a Revision of Young's Manual of Astronomy, Vol. I, The Solar System," Boston: Ginn and Co., 1945.
101. VICTOR, W. K., STEVENS, R., and GOLOMB, S. W. (Ed.), "Radar Exploration of Venus: Goldstone Observatory Report for March–May 1961," Calif. Inst. Tech., Jet Propulsion Lab. Tech. Rept. No. 32-132, August 1, 1961.
102. ASHBROOK, J., "A New Determination of the Rotation Period of the Planet Mars," *Astronom. J.*, **58**(1210): 145–155, August 1953.
103. PEEK, B. M., "The Planet Jupiter," New York: Macmillan, 1958.
104. CAMICHEL, H., "Mesure de la Durée de Rotation de Saturne d'Après une Tache," *Bulletin Astronomique* **20**(2): 141–144, 1956.
105. GÜNTHER, O., "Der Rotationslichtwechsel des Neptun," *Astronomische Nachrichten*, **282**(1): 1–14, 1955.
106. WALKER, M. F., and HARDIE, R., "A Photometric Determination of the Rotational Period of Pluto," *Publ. Astronom. Soc. Pacific*, **67**(397): 224–231, August 1955.
107. MAYER, C. H., "Radio Emission of the Moon and Planets," in: *Planets and Satellites*, ed. by G. P. Kuiper and B. M. Middlehurst, Chicago: University of Chicago Press, 1961, pp. 442–472.

Appendix A

DETAILED ANALYSIS OF THE FREE OSCILLATIONS

ELASTIC VIBRATIONS OF A SPHERE

At frequencies for which the spectrum is recognizably discrete, the normal modes involve a major part of the earth. The characteristic length associated with these vibrations is large compared with the inhomogeneities of the continent-ocean system. Hence, it is customary to assume spherical symmetry for the earth. The earth is then defined by the radial functions determining the initial distribution of unperturbed density, $\rho(r)$, and the Lamé constants, $\mu(r)$ and $\lambda(r)$.

In the analysis of the free vibrations of the solid earth we first describe the equilibrium state and then superimpose on the static solution a perturbation velocity field v_i controlled by the elastic gravitational and rotational restoring forces. As we shall see, the rotational restoring forces can be treated as a perturbation, but for the low frequency vibrations the effect of gravity must be explicitly taken into account.

Consider a rotating isotropic solid. The equilibrium conditions, referred to a coordinate axis rotating at angular velocity Ω_j, are

$$\left.\begin{aligned} \text{grad } U &= -\frac{1}{\rho}\,\text{grad } p\ , \\ U &= U_g(r) - \frac{1}{2}\,r^2\,\Omega^2\cos^2\theta\ , \end{aligned}\right\} \qquad \text{(A1)}$$

where p is the hydrostatic pressure and ρ is the density. The potential U is made up of two parts. $U_g(r)$ is the spherically symmetrical potential of the gravitational forces satisfying Poisson's equation,

$$\nabla^2 U_g = 4\pi G\rho\ . \qquad \text{(A2)}$$

The second term in the equation for the total potential U is proportional to the square of the angular velocity Ω, and represents the cylindrically symmetrical centrifugal forces; r is the distance from the earth's center; θ is the latitude.

The surfaces, $U =$ constant, are ellipsoids of revolution with the maximum ellipticity at the surface. An approximate representation of the equilibrium surface is

$$r = a\left[1 + \epsilon_T\left(\frac{1}{3} - \sin^2\theta\right)\right], \qquad \text{(A3)}$$

where ϵ_T is the hydrostatic value for the flattening of the ellipsoid. The theoretical value for the flattening of a body rotating at the angular velocity of the earth and also having the internal density distribution of the earth is 1/299.8. The observed value is 1/298.2 (Section 2). Thus the actual surface approximates the surface of hydrostatic equilibrium to within 2 parts in 300, and the surfaces of equilibrium deviate from spherical surfaces by only one part in 300 throughout the earth.

The free oscillations are studied by superimposing a small stress on the hydrostatic configuration described by equation A1.[1,2] The total stress tensor, p_{ij}, at any point within the earth is the sum of a hydrostatic pressure $p\delta_{ij}$ and τ_{ij}, an elastic stress tensor measured from the initially compressed equilibrium state:

$$p_{ij} = -p\,\delta_{ij} + \tau_{ij}\ . \qquad \text{(A4)}$$

It is assumed that τ_{ij} is linearly related to the strain tensor ϵ_{ij} through the local elastic parameters. At any point the divergence of the elastic stress tensor is

$$\frac{\partial \tau_{ij}}{\partial x_j} = (\lambda + \mu)\frac{\partial^2 u_j}{\partial x_i\,\partial x_j} + \mu\frac{\partial^2 u_i}{\partial x_j^{\,2}}\ , \qquad \text{(A5)}$$

where λ and μ are the Lamé constants. The element u_1 is the displacement vector measured from the initial equilibrium stress state.

The equations of motion referred to a coordinate

system fixed to the mean body are

$$\rho \frac{\partial^2 u_i}{\partial t^2} + 2\rho \epsilon_{ijk} \Omega_j \frac{\partial u_k}{\partial t} = -\rho \frac{\partial U}{\partial x_i} + \frac{\partial p_{ij}}{\partial x_j} , \tag{A6}$$

provided that the ellipticity and the quadratic terms in the displacement are neglected. Neglecting the ellipticity implies that the initial pressure distribution, p_0, and initial density distribution, ρ_0, are spherically symmetrical functions of the radius. In addition, we will assume that the elastic constants are radial functions. The assumed spherical symmetry of the earth neglects the near-surface continent-ocean structure. The supposition is that the inhomogeneities of the near-surface material do not extend to great depths. The low frequency vibrations involve the earth as a whole and are not affected in any major way by surface inhomogeneities.

The equations of motion referred to a rotating system differ from the nonrotating case because of two terms involving the angular velocity. The total potential U includes a term varying as the square of the angular velocity. The smallness of the cylindrically symmetrical centrifugal force terms follows from

$$\left.\begin{aligned} \frac{\partial^2 U}{\partial x_i^2} &= 4\pi\epsilon\rho - 2\Omega^2 , \\ \frac{2\Omega^2}{4\pi\epsilon\rho} &= \frac{1}{300} . \end{aligned}\right\} \tag{A7}$$

The term involving the angular velocity of the left-hand side of equation A6 is the Coriolos or gyroscopic force. Its effect is to couple motion along the various coordinate axes since the rotation deflects the oscillating particles. The rotational coupling introduces a complexity in the motion that is important in interpreting the detailed observations of the earth's free vibrations.

Assume a time variation of the displacement of

$$u_k = q_k e^{i\omega t} , \tag{A8}$$

where q_k is a complex amplitude and ω, the angular frequency of vibration, assumes only discrete values. In terms of the complex amplitude the equations of motion are

$$-\omega^2 \rho q_k + 2i\rho\omega\epsilon_{krs} \Omega_r q_s = -\rho \frac{\partial U}{\partial x_k} + \frac{\partial p_{kj}}{\partial x_j} . \tag{A9}$$

The gyroscopic forces couple the motion along the coordinate axes and, in spherical coordinates, equation A9 is not separable. However, the strength of the coupling depends upon the ratio of the angular velocity to the frequency of oscillation. A rough estimate of the period of the fundamental oscillation is provided by the ratio of the circumference to the average elastic wave velocity,

$$\frac{2\pi a}{\bar{V}} = \frac{2\pi(6.3 \times 10^8)}{8 \times 10^5} = 4 \times 10^3 \text{ sec} . \tag{A10}$$

Thus the lowest order oscillation should have a period on the order of an hour. This frequency is an order of magnitude greater than 2Ω and we may suppose that the effect of the earth's rotation can be treated as a perturbation of the solution for a nonrotating earth. Let

$$\delta = \frac{\Omega}{\omega^0} , \tag{A11}$$

where ω_0 is the frequency of oscillation in the absence of rotation. The perturbation expansions are

$$\left.\begin{aligned} q_k &= q_k^0 + \delta q_k^1 + \cdots , \\ \omega &= \omega^0 (1 + \alpha\delta + \cdots) , \end{aligned}\right\} \tag{A12}$$

where q_k^0 is the complex amplitude in the absence of rotation. Introducing equations A12 into equation A9 we note that the gyroscopic terms are of order δ whereas the centrifugal terms are of order δ^2. The term α in equations A12 determines the magnitude of the frequency shift resulting from rotation. The equations of motion to zero and

first order in δ are:

$$\left.\begin{aligned} -\rho(\omega^0)^2 q_k^0 + \rho\frac{\partial U^0}{\partial x_k} - \frac{\partial p_{kj}{}^0}{\partial x_j} &= 0 , \\ -\rho(\omega^0)^2 q_k^1 + \rho\frac{\partial U^1}{\partial x_k} - \frac{\partial p_{kj}}{\partial x_j} &= -2\rho(\omega^0)^2 \alpha q_k^0 + 2\rho i(\omega^0)^2 \epsilon_{krs} z_r q_s^0 , \\ \Omega_j &= \Omega z_j . \end{aligned}\right\} \quad \text{(A13)}$$

The motion to zero order in δ is governed by elasticity, by the gravitational forces arising from the alteration of the density field by the motion, and by the attraction of the deformed interfaces. Equations A13 and A2 determine the displacement and the gravitational potential. These solutions are subject to the conditions that: (1) The solution is regular at the origin. (2) The stresses vanish on the deformed surface of the earth. (3) The internal and external gravitational potentials and their gradients are equal at the deformed external surface. In addition, the stress components must be continuous at any internal surface of discontinuity, as must the gradient of the gravitational potential.

The equations of motion are separable in spherical coordinates if rotation is neglected. A solution with spherical symmetry is:

$$\left.\begin{aligned} q_j &= \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\left[\sqrt{l(l+1)}\,C_j^{m}{}_{,l}(\theta,\lambda)\,{}_nT_l(r)\right. \\ &\quad \left. + P_j^{m}{}_{,l}(\theta,\lambda)\,{}_nW_l(r) + \sqrt{l(l+1)}\,B_j^{m}{}_{,l}(\theta,\lambda)\,{}_nV_l(r)\right] , \\ U(r) &= \sum_{l=0}^{\infty}\sum_{m=-l}^{l} X_l^{m}(\theta,\lambda)\,{}_n\psi_l(r) , \end{aligned}\right\} \quad \text{(A14)}$$

where the vector spherical harmonics C_i, B_i, and P_i are defined by[3]

$$\left.\begin{aligned} C_i^{m}{}_{,l}(\theta,\lambda) &= \frac{1}{\sqrt{l(l+1)}}\epsilon_{ijk}\frac{\partial}{\partial x_k}(x_k X_l^{m}) , \\ P_i^{m}{}_{,l}(\theta,\lambda) &= \frac{x_i}{r} X_l^{m} , \\ B_i^{m}{}_{,l}(\theta,\lambda) &= -\frac{1}{\sqrt{l(l+1)}} + \frac{\partial}{\partial x_i} X_l^{m} . \end{aligned}\right\} \quad \text{(A15)}$$

X_l^m denotes the complex spherical harmonics. T, W, V, ψ are radial functions and the subscript n denotes the overtone or the number of radial zeros. It should be mentioned that in the notation adopted the letters l and m are reserved for the angular order of the spherical harmonics and are associated with the latitude θ and longitude λ, respectively. The coordinate subscripts are separated from the order subscripts by commas.

The vectors C_i, P_i, and B_i are mutually orthog-

onal. Motions described by

$$q_i^0 = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \sqrt{l(l+1)}\, C_i^m{}_{,l}(\theta, \lambda)\, {}_nT_l(r) \qquad \text{(A16)}$$

involve no radial component; the particles remain on a spherical surface throughout an oscillation. The density remains constant since

$$\frac{\partial q_i^0}{\partial x_i} = 0 . \qquad \text{(A17)}$$

These oscillations are termed the toroidal oscillations. They are particularly simple since they do not involve coupling with the density field, and, as a result, the gravitational field remains constant.

The motions described by P_i and B_i are characterized by the vanishing of the radial component of the curl of the displacement. They are labeled *spheroidal* oscillations. They involve both radial and horizontal components of the displacement; therefore the gravitational forces come into play.

NUMERICAL METHODS FOR OBTAINING THE EIGEN FREQUENCIES

Two methods have been used to obtain the eigen frequencies for an earth with a radial variation of its elastic properties. In an approach taken from Pekeris the equations are reduced to a set of first-order ordinary differential equations which are integrated by the Runge–Kutta method. In an alternate scheme, matrix methods are used to join up solutions valid in homogeneous layers.

In order to illustrate the methods that have been applied, we will consider the simpler case of *toroidal* oscillations. The equation of motion for toroidal oscillation is

$$\mu\left(\frac{d^2T}{dr^2} + \frac{2}{r}\frac{dT}{dr}\right) + \frac{d\mu}{dr}\left(\frac{dT}{dr} - \frac{T}{r}\right) + \left[(\omega^0)^2 \rho^0 - \frac{l(l+1)}{r^2}\mu\right] T = 0 , \qquad \text{(A18)}$$

where the subscripts of the radial function T have been omitted. The boundary conditions require that the stress should vanish at the outer surface of the elastic mantle and, since the core is fluid, should also vanish at the inner surface.

Alterman, Jarosch, and Pekeris[4] formally obviated the need to evaluate the empirically obtained derivative of rigidity by introducing the variables

$$\left.\begin{aligned} y_1 &= T , \\ y_2 &= \mu\left(\frac{\partial T}{\partial r} - \frac{T}{r}\right) . \end{aligned}\right\} \qquad \text{(A19)}$$

Equation A18 is then equivalent to the system

$$\left.\begin{aligned} \frac{dy_1}{dr} &= \frac{1}{r} y_1 + \frac{1}{\mu} y_2 , \\ \frac{dy_2}{dr} &= \left[\frac{\mu(l^2 + l - 2)}{r^2} - (\omega^0)^2 \rho^0\right] y_1 - \frac{3}{r} y^2 , \end{aligned}\right\} \qquad \text{(A20)}$$

where the boundary conditions are that y_2 must vanish at the outer and inner boundaries of the earth's mantle. Equations A20 can be integrated by the Runge–Kutta method and the eigen frequencies determined as a function of the integer l. The analysis of the spheroidal oscillations is similar, but it involves the simultaneous solution of six first-order equations.

An alternative method, used by Gilbert and MacDonald,[5] employs the known solution for a homogeneous spherical shell. The earth is assumed to be a number of concentric spherical elastic shells. The solution for each shell is continued to the next shell by use of the appropriate boundary condition.

The displacement in a toroidal oscillation for a homogeneous elastic shell is

$$g_i^0 = \sum_{l=1}^{\infty} \sum_{m=-l}^{l} \sqrt{l(l+1)}\, C_{i,l}^{m}(\theta, \lambda) \left[a_l^m\, j_l(kr) + b_l^m\, y_l(kr)\right], \tag{A21}$$

where

$$k = \frac{\omega^0}{V_s} \tag{A22}$$

is the wave number. The values j_l and y_l are spherical Bessel functions of the first and second kind. The coefficients a and b are determined by the boundary conditions that the radial traction vanish, i.e.,

$$\sum_{l=1}^{\infty} \sum_{m=-l}^{l} \sqrt{l(l+1)}\, C_{i,l}^{m}(\theta, \lambda) \cdot \cdot \left\{\mu k a_l^m \left[j_l'(kr) - (kr)^{-1} j_l(kr)\right] + \mu k b_l^m \left[y_l'(kr) - (kr)^{-1} j_l(kr)\right]\right\} = 0, \tag{A23}$$

at the inner and outer surfaces. If we denote the bracketed terms in equations A21 and A23 by R and S respectively, than

$$\begin{bmatrix} R \\ S \end{bmatrix} = \begin{bmatrix} j_l(kr) & y_l(kr) \\ \mu k\left[j_l'(kr) - (kr)^{-1} j_l(kr)\right] & \mu k\left[y_l'(kr) - (kr)^{-1} y_l(kr)\right] \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}. \tag{A24}$$

Let s denote the s spherical shell; r_{s+1} identify the bottom and top radii of the s spherical shell, respectively. Equations A24 take the form

$$\begin{bmatrix} R_s(r) \\ S_s(r) \end{bmatrix} = \mathcal{A}_s(r) \begin{bmatrix} a_s \\ b_s \end{bmatrix}, \tag{A25}$$

where $\mathcal{A}$ is a 2,2 matrix. $\mathcal{A}_s$ is a function of the density and rigidity of the layer whose inner radius is r_s. Equation A25 can be written

$$\begin{bmatrix} a_s \\ b_s \end{bmatrix} = \mathcal{A}_s^{-1}(r) \begin{bmatrix} R_s(r) \\ S_s(r) \end{bmatrix}, \tag{A26}$$

provided $\mathcal{A}_s$ is not singular. By combining equations A25 and A26 we relate the stress and displacement at the level r_s to the same quantities at r_{s+1}:

$$\begin{bmatrix} R_s(r_s) \\ S_s(r_s) \end{bmatrix} = \mathcal{A}_s(r_s)\, \mathcal{A}_s^{-1}(r_{s+1}) \begin{bmatrix} R_s(r_{s+1}) \\ S_s(r_{s+1}) \end{bmatrix}. \tag{A27}$$

In terms of the inner, r_I, and outer, r_0, radii of the mantle,

$$\begin{bmatrix} R_I(r_I) \\ S_I(r_I) \end{bmatrix} = \mathcal{C} \begin{bmatrix} R_N(r_0) \\ S_N(r_0) \end{bmatrix}, \tag{A28}$$

where

$$\left.\begin{aligned} \mathcal{C} &= \mathcal{B}_1\, \mathcal{B}_2 \cdots \mathcal{B}_N, \\ \mathcal{B}_s &= \mathcal{A}_s(r_s)\, \mathcal{A}_s^{-1}(r_{s+1}), \end{aligned}\right\} \tag{A29}$$

The stress must vanish at both the inner and outer surfaces. Therefore equations A28 become

$$\left.\begin{aligned} \begin{bmatrix} R_I(r_I) \\ 0 \end{bmatrix} &= \mathcal{C} \begin{bmatrix} R_N(r_0) \\ 0 \end{bmatrix}, \\ \mathcal{C}_{21} &= 0. \end{aligned}\right\} \tag{A30}$$

The roots of equations A30 are, then, resonant frequencies for the toroidal oscillations.

Spheroidal oscillations may be treated similarly. The matrices are 6×6 and the roots are deter-

TABLE A1.—*Rotational Splitting of Toroidal and Spheroidal Oscillations* ($l=2$)

m	Period (min.)	
	Toroidal	Spheroidal
2	41.9	52.1
1	42.1	52.9
0	42.3	53.7
-1	42.5	54.5
-2	42.7	55.3

mined by the vanishing of a partition of the matrix. In this method the elements of the matrices and the inversions are evaluated numerically.

EFFECT OF ROTATION ON THE EIGEN FREQUENCIES

The numerical calculations of the previous section can be used to obtain the eigen frequencies for the nonrotating earth. For this case the solution is degenerate with respect to the ordinal number m (see equations A14); the two possible signs of m are symmetrical for a stationary sphere. Rotation removes the degeneracy with respect to the azimuthal number m. Each line in the spectrum is split into a multiplet of $2l+1$ lines where the frequencies are given by

$$\omega_l^m = \omega_l^0 + m\sigma(l)\Omega \ . \qquad \text{(A31)}$$

A perturbation calculation[6–8] shows that for toroidal oscillations σ is

$$\sigma(l) = \frac{1}{l(l+1)} \ , \qquad \text{(A32)}$$

whereas in the case of spheroidal oscillations,

$$\sigma(l) = \frac{\int_0^{r_0} \rho r^2 (2WV + V^2) dr}{\int_0^{r_0} \rho r^2 \left[W^2 + l(l+1)V^2\right] dr} \ , \qquad \text{(A33)}$$

where r_0 is the outer radius. W and V are the radial functions in equations A14. Table A1 illustrates the splitting of the spheroidal and toroidal oscillations for the Gutenberg model earth.

REFERENCES

1. PEKERIS, C. L., ALTERMAN, Z., and JAROSCH, H., "Rotational Multiplets in the Spectrum of the Earth," *Phys. Rev.*, **122**(6): 1692–1700, June 15, 1961.
2. MACDONALD, G. J. F., and NESS, N. F., "A Study of the Free Oscillations of the Earth," NASA Technical Report R–136, 1962; also *J. Geophys. Res.*, **66**(6): 1865–1911, June 1961.
3. MORSE, P. M., and FESHBACH, H., "Methods of Theoretical Physics," New York: McGraw-Hill, 1953.
4. ALTERMAN, Z., JAROSCH, H., and PEKERIS, C. L., "Oscillations of the Earth," *Proc. Roy. Soc. London*, **252A**(1268): 80–95, August 25, 1959.
5. GILBERT, F., and MACDONALD, G. J. F., "Free Oscillations of the Earth: I. Toroidal Oscillations," *J. Geophys. Res.*, **65**(2): 675–693, February 1960.
6. BACKUS, G., and GILBERT, F., "The Rotational Splitting of the Free Oscillations of the Earth," *Proc. Nat. Acad. Sci.*, **47**(3): 362–371, March 1961.
7. PEKERIS, C. L., ALTERMAN, Z., and JAROSCH, H., "Rotational Multiplets in the Spectrum of the Earth," *Phys. Rev.*, **122**(6): 1692–1700. June 15, 1961.
8. MACDONALD, G. J. F., and NESS, N. F., "A Study of the Free Oscillations of the Earth," NASA Technical Report R–136, 1962; also *J. Geophys. Res.*, **66**(6): 1865–1911, June 1961.

EVIDENCE FROM THE MOON'S SURFACE FEATURES FOR THE PRODUCTION OF LUNAR GRANITES

JOHN A. O'KEEFE

AND

WINIFRED SAWTELL CAMERON

Goddard Space Flight Center

If tektites come from the moon, they indicate a granitic constitution for a considerable portion of its outer crust, especially if allowance is made for the chemical effects of volatization (loss of alkalies). The evidence for granitic rocks from the morphology of the moon's surface is considered in this paper. The displacement of the center of the moon's visible surface with respect to the center of mass is considered evidence of isostasy on the moon. It is shown here that this displacement is not merely a limb phenomenon, and is probably not explicable in terms of the kind of umpy interior proposed by Urey to explain the differences of the moments of inertia. Evidence for characteristic granite topographic features, including tholoids and laccoliths, is summarized. The possibility that the maria were produced by the fluidization of volcanic ash is considered. The transporting fluid is considered to be the contained gases as in a terrestrial ash flow. Ash flows on the earth are responsible for the emplacement of a large fraction of all volcanic ash. From a study of the physics of ash flows, it is determined that the fluidization should be even more effective on the moon. It is concluded that the morphology of the moon's surface does not contradict the notion that large amounts of acid rock are present on its surface.

INTRODUCTION

Evidence has been adduced from the writing of Nininger, Dietz, O'Keefe, Varsavsky, Gold, and Chapman[1–6] in support of the hypothesis, originally enunciated by Verbeek,[7] that tektites come from the moon. Chapman's work in particular has shown that the external features of the australites indicate a velocity of entry into the earth's atmosphere between 10 and 13 km per second. Taken in conjunction with the observed surface distribution, this figure excludes direct origin from the earth, and it likewise excludes normal meteoric orbits. It is, on the other hand, entirely consistent with the theory of origin from the moon. Since particles coming from the moon as a result of meteorite impacts are to be expected a priori, and since these ought to be different from meteorites, it is highly likely that at least some of these particles are tektites.

Is the hypothesis of a lunar origin for tektites consistent with what is known about the structure of the moon's surface?

In the present paper we will attempt to relate the observable features of the moon's surface with the kind of rocks which would be expected to yield tektites. First an effort will be made to deduce the type of rock which formed the parent of the tektites; then the common surface features of the earth associated with rock of this kind will be noted; and finally we shall attempt to identify these features on the moon.

THE ROCK WHICH YIELDS TEKTITES

In the first place, it is clear that, if they are lunar, the tektites are related to the igneous rocks since sedimentary processes in the ordinary sense do not occur on the moon—unless Gilvarry's radical hypothesis of lunar oceans is correct.[8] In the second place, tektites belong with the acid igneous rocks of the earth (those with a silica

content of 65 percent or more) rather than with the intermediate rocks (55 to 65 percent silica) or the basic rocks (55 percent or less silica). No known tektite has less than 65 percent silica; nearly all have more than 68 percent.[9]

Tuttle and Bowen showed that the majority of the acid igneous rocks can be thought of as consisting principally of three minerals—quartz, orthoclase, and albite—in roughly equal proportions.[10]

Petrologists express such a relation with a triangular diagram, such as that in figure 1, in which each corner is labeled with the name of one of the three constitutents. The relative proportions of the three constituents are expressed by the position of a point on the diagram; its distance from any side of the triangle is proportional to the abundance of the mineral marked on the opposite vertex. The normative composition is the theoretical equilibrium mineral composition as computed from chemical analysis of the rock.

It can be seen that granites and their volcanic equivalents, rhyolites, cluster around the center of the triangular diagram (figure 1). There is experimental evidence for this behavior and it can be explained theoretically. If the composition of the first fraction of liquid produced by melting in this system—in the first instance, at atmospheric pressure—is projected onto this diagram, it is found that the granites occupy the locus of lowest melting point, near the center of the diagram. At higher water pressures there is some migration of the low-melting-point trough, but it remains near the center. The granites therefore are believed to be formed either as the last portion of a magma to solidify (magmatic differentiation) or the first portion of a solid rock to liquefy. In the first case, they might represent the last remaining liquid in a large underground reservoir of magma; in the latter case they might represent a liquid sweated out of a solid mass at high temperature. These possibilities are further discussed elsewhere by Lowman.[12]

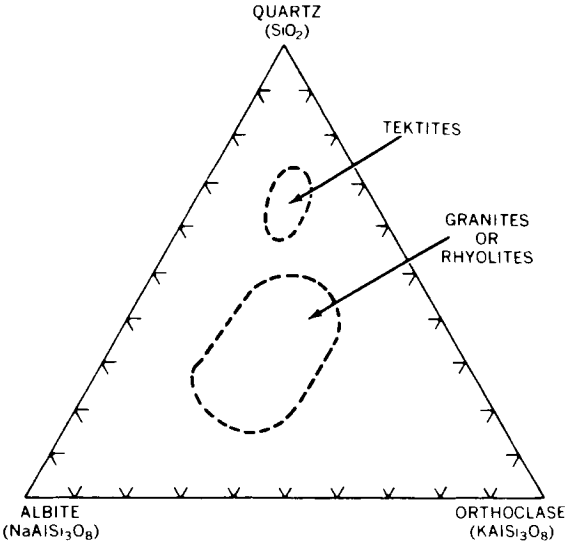


FIGURE 1.—Normative compositions of granites and tektites, weight percentage (modified from reference 10; tektite norms from reference 11).

The important point to see here is that tektites as they are found do not fit the picture. As is shown in figure 1, tektites are on the quartz side of the granite field (and the low-melting-point trough), in a region where very few igneous rocks are located. If the tektites fell below the central region (i.e., on the orthoclase-albite side), it would be possible to explain their composition as due to incompleteness of one of the igneous processes mentioned. There are sound reasons for thinking that, either on the earth or on the moon, many acid igneous rocks have been derived from basaltic magma or from some rock of lower silica content. Terrestrial rocks that would be above the trough in figure 1 can be explained as remelted sandstones, since sand is mostly silica; but this explanation is not available for lunar materials. The problem is so serious that many students of tektites, including Urey, reject the theory of a lunar origin of tektites, for just this reason, emphasizing that the chemistry of silicates ought to be the same on the moon as on the earth.

A possible solution of the problem has been provided by Cohen[13] and Taylor,[14] who have pointed out the significance of volatilization in altering the composition of tektites. Taylor showed that the flanges of australites, which have been exposed to two meltings, are about 20 percent poorer in soda and potash than the cores, which have been exposed to only one melting. Cohen showed that the ratio of the volatile element germanium to the less volatile gallium was lower in the flanges than in the cores. Taylor applied his work only to the australite problem, but Cohen

made the interesting suggestion that the same process had been at work in the formation of the taktites in general.

Cohen's hypothesis is in good accord with many other physical and chemical properties of tektites, namely:

1. The remarkable lack of water and other volatiles;[15]
2. The presence of FeO rather than Fe_2O_3;
3. The almost total lack of magnetization;[16]
4. The glassy nature of the tektites themselves.

These characteristics by themselves suggest a thorough heating at a temperature above that of ordinary lavas. At high temperatures, the escape of alkalies is commonly observed in the laboratory (reference 10, p. 6).

It is therefore plausible to support that the tektites are derived from a material much like a granite or other acid igneous rock, provided that we accept the idea of a period of heating at very high temperatures, lasting several minutes, during some part of the history of the tektites. Perhaps the heating occurred during a hypothetical formation of the tektites as ablation drops in the atmosphere. The mechanism of ablation heating is known, from the example of the australite flanges, to be competent to account for the volatilization. An alternative possibility, that the heating resulted from impact on a surface such as that of the moon, encounters the difficulty that the time duration of impact heating appears to be short, as judged from the lack of homogeneity of terrestrial impact glasses.

Thus the first part of this discussion is completed; we find that the tektites belong with the granitic kindred. It should be mentioned that the term "granite" is used here primarily in a chemical sense; strictly speaking, granite is a coarse-grained rock, usually formed at a great depth and exposed later by erosion. Although there may be some granite in the moon which has been exposed in craters, it should be uncommon; the granitic magmas will typically have formed fine-grained volcanic rocks such as rhyolite and tuff.

SURFACE FEATURES ASSOCIATED WITH GRANITIC ROCKS

The second part of this discussion is the listing of land forms that are associated with granitic rocks, especially those which might be visible on the moon. They are:

1. Isostasy—continental blocks of crust supported in a more or less hydrostatic fashion by flotation above a heavier substratum. The continents of the earth are made essentially of granite or its derivatives. The ocean basins, however, are apparently underlain by basaltic rock with a density about 10 percent higher.
2. Batholiths—large masses of granitic rock typically found in the cores of large folded mountain belts such as the Appalachians. Since there are no folded mountains of this type (consisting of thick accumulations of sedimentary rock) on the moon, we do not expect evidence that typical batholiths exist there.
3. Laccoliths—mushroom-shaped bodies, usually composed of acid igneous rock, intruded between the layers of a stratified country rock and causing the overlying strata to become dome-shaped.
4. Peleean volcanism, characterized by:

a. Extrusion of lava in the form of spines and ridges, owing to high veiscosity.

b. Production of ash, both as ordinary ash falls and especially as ash flows, in which the solid material is fluidized and transported by the contained gases.

EVIDENCE FOR THE EXISTENCE OF ISOSTASY, LACCOLITHS, AND PELEEAN VOLCANISM ON THE MOON

Isostasy

The fundamental evidence for isostasy on the moon is the difference of position between the moon's center of gravity and its center of figure. The difference is well established; for example, it has been regularly employed for the calculation of the eclipses of the sun (see the *American Ephemeris and Nautical Almanac for 1962*, p. 291). It amounts to approximately 1 km, in the sense that the center of figure is south of the center of gravity. The basic principle is that observations of the moon give the position of the center of the moon's visible face, whereas theory refers to the center of gravity. A discrepancy appears between theory and observation, by which the observed position is persistently south of the calculated one.

The discrepancy cannot be removed by amending the lunar theory, since this does not contain constant terms in the celestial latitude. It is therefore attributed to a difference between the centers. Similar differences might in principle exist in the east-west direction or in the radial direction; but there is no easy observational way to find them.

The reality of the effect has been questioned on the ground that the moon's southern limb is conspicuously rougher than the northern limb (private communication with H. C. Urey). Hence the effect might be due merely to the fact that at the limb we are looking at the ridge lines. In rough country their average height is higher than the ground as a whole.

It is possible, however, to verify that the same displacement exists to an even greater extent for points well away from the limb. The reader may examine the maps of Baldwin[17] or the recent U.S. Army Map Service Topographic Map of the Moon, produced under the supervision of A. L. Nowicki. Both give elevations referred to a center which fits the limb observations, and hence lies south of the center of gravity. Nevertheless, it can be seen at once that the average height of the ground in the southern portion is greater, on these charts, then in the northern part. This means that the center of figure as derived from the maps falls even further south of the center of gravity than the center as derived from limb observations.

For obtaining a quantitative estimate, use was made of the catalog of 150 points, on the surface of the moon, established by G. Schrutka-Rechtenstamm.[18] His catalog gives rectangular coordinates in the usual system, the unit being the radius of the moon, referred to the following coordinates for the center of the crater Mösting A:

$\xi = -0.08992$ (ξ positive to west),

$\eta = -0.05551$ (η positive to north),

$\zeta = +0.99521$ (ζ positive toward the earth).

These coordinates refer to a center derived from limb observations which is, as usual, about one km south of the center of gravity.

At least squares solution was made for the center of that sphere which most nearly fits the coordinates given by Schrutka-Rechtenstamm. Since the solution is almost independent of the correction in the radial (ζ) direction, the ζ correction was arbitrarily fixed at zero. The derived corrections to the ξ and η coordinates are $\Delta\xi = -0.4 \pm 0.4$ (m.e.) and $\Delta\eta = +1.4 \pm 0.4$ (m.e.), from which it follows that the center of figure should be a kilometer or so farther south. The correction in the east–west direction is small and uncertain. This result tends to confirm what is seen on the maps, as would be expected, since the maps are based on the coordinates of Schrutka-Rechtenstamm and his predecessors. The result is not, of itself, as decisive as it might appear, since the coordinates available omit the outer rim of the moon's disk. Owing to foreshortening, the outer rim forms only a small fraction of the visible *disk*, although it makes up nearly half of the visible *hemisphere*. A significant point is that the rough continental areas are found to be, in fact, higher than the smooth maria, both in the coordinate calculations and on the maps.

On the whole, however, it appears impossible to doubt the reality of the displacement of the centers. Urey's contrary statement (reference 19, pp. 22–25) is due to the fact that the occultation data which he used were calculated from the center of the moon's face rather than from the center of gravity.

Two explanations might exist for the displacement of the center of figure.

1. It might be due to isostasy. On the earth the center of mass is at the center of figure of the *geoid* (sea level surface) according to a theorem in potential theory. Since the Pacific hemisphere, which is centered near 180° longitude and 30°S latitude, is nearly all water, whereas the other hemisphere is nearly all land, it is clear that for the earth there is also a displacement of the center of figure with respect to the center of mass. It is well known that this displacement exists because the continents are essentially granite, whereas the ocean bottoms are essentially basalt, so that there is an excess of density in the ocean hemisphere. The displacement of the moon's center of figure may be due to a similar structure. A division of the moon into roughly two physiographic parts is apparent to the eye and telescope from coloration and comparative elevation; the lower elevation is associated usually with darker colored material than the higher elevation. This suggests an excess of density in the maria.

2. A second possibility, to which Urey, Elsasser, and Rochester[20] have drawn attention, is that the moon is inhomogeneous in a more or less random way, and consists of a finite number of lumps of unequal density.

To distinguish between these two hypotheses, we make use of the evidence from the second harmonics of the moon's gravitational field which indicates density inhomogeneities. If the irregularities in density are at the surface, and are supported in a more or less hydrostatic way, then the net effect on the external gravitational field will be small, since, by Archimedes' principle, the floating continental blocks will displace their own weight of basic rock, leaving the gravitational field approximately the same, at least at great distances. If, on the other hand, the inhomogeneities are irregularly distributed through the mass of the moon, then much larger gravitational effects are to be expected.

The external gravitational potential U of any planet can be expressed as a series of spherical harmonics, of the form

$$U = \sum_{n=0}^{\infty} \frac{A_n}{r^{n+1}} S_n(b, l) , \qquad (1)$$

where r is the radius, b the latitude, l the longitude; the values of $S_n(b, l)$ are surface spherical harmonics, and the A_n's are coefficients of the harmonics. In this expression, the harmonic of zero order represents the field which the body would have if the mass were concentrated at one point. If the origin of coordinates is taken at the center of mass, then by a theorem in potential theory, the three first harmonics in the potential vanish; this is equivalent to the above remark about the center of figure of the geoid coinciding with the center of mass of the earth.

Since the effect of irregularities in the moon's gravitational field must be measured from the action of the irregularities on the earth or the sun (together with the reaction on the moon), it follows that the effects will die out rapidly as we use higher and higher harmonics. The second harmonics are the lowest harmonics whose effects can be expected to reflect the irregularities of the moon; they are also the highest harmonics which can now be detected.

We wish to compare the observed variations in the second harmonics over the moon's surface with the variations to be expected from a theory of lumps. We shall first obtain an expression for the moment of inertia, I, in an arbitrary direction, in terms of quantities determined in libration theory. Next we shall find the deviation of I from its average value I^0, square the deviation, and integrate it over the sphere, to find the mean squared value. This we shall compare with the mean squared deviation to be expected from a statistical analysis like that of Urey, Elsasser, and Rochester,[20] but more rigorous. Finally, we shall compare the lumpiness needed to produce these effects in the second harmonic with those required to produce the observed effect in the displacement of the centers, and thus see whether the gravitational field is notably smoothed, as would be expected in the case of isostasy.

It is also a general result of potential theory that the coefficients of the spherical harmonics of order n are proportional to linear combinations of the nth moments of the mass distribution. Second harmonics in particular are functions of the moments of inertia. The theory is given in standard texts—in reference 21, for example. If the moments of inertia around the ξ, η, and ζ axes are respectively denoted by B', C', and A', then the theory of the moon's figure gives the quantities

$$\left.\begin{aligned} L' &= \frac{3}{2}\frac{C' - A'}{M'a'^2} = +0.0003734, \\ K' &= \frac{3}{2}\frac{B' - A'}{M'a'^2} = +0.000070, \end{aligned}\right\} \qquad (2)$$

where, as in reference 21, primes refer to the moon. Also,

$$J' = \frac{3}{2}\frac{C' - \frac{1}{2}A' - \frac{1}{2}B'}{M'a'^2}$$

$$= L' - \frac{1}{2}K' = +0.0003384. \tag{3}$$

The general expression for the moment of inertia I around an axis with direction cosines λ, μ, and ν, referred to the axes of A', B', and C', respectively, is

$$I = A'\lambda^2 + B'\mu^2 + C'\nu^2 . \tag{4}$$

Setting

$$\lambda = \cos b \cos l ,$$

$$\mu = \cos b \sin l ,$$

$$\nu = \sin b ,$$

where b is the selenographic latitude, and l the selenographic longitude, we have

$$I = A' \cos^2 b \cos^2 l + B' \cos^2 b \sin^2 l + C' \sin^2 b . \tag{5}$$

Following Jeffreys' treatment of a similar problem (reference 21, p. 133), we look for a linear expression for I in terms of both the mean value

$$I_0 = \frac{A' + B' + C'}{3} , \tag{6}$$

and the spherical harmonics (in Jeffreys' notation)

$$P_2^0 = \frac{3}{2}\sin^2 b - \frac{1}{2} , \tag{7}$$

and

$$P_2^2 \cos 2l = \frac{3}{2}\cos^2 b \cos 2l , \tag{8}$$

with J' and K' in the coefficients of the spherical harmonics. We find

$$I = I_0 + M'a'^2 \left(\frac{4}{9}J'P_2^0 - \frac{2}{9}K'P_2^2 \cos 2l\right) , \tag{9}$$

as may be verified by evaluation at the moon's poles ($b = \pm 90°$, $I = C'$), at the center of its visible face ($b = 0$, $l = 0$, $I = A'$), and at the east and west limbs ($b = 0$, $l = \pm 90°$, $I = B'$). These three conditions suffice to fix the three coefficients of the linear form.

The mean squared value of $(I - I_0)$ over the sphere Ω is then

$$\overline{\Delta I^2} = \overline{(I - I_0)^2} = \frac{1}{4\pi}\iint_{\Omega} (\Delta I)^2 d\omega , \tag{10}$$

where $d\omega$ is the surface element of the sphere.

In the integration, the surface integral of the cross products of the spherical harmonics vanishes, since spherical harmonics are orthogonal: hence

$$\overline{(\Delta I)^2} = \frac{1}{4\pi}\iint_{\Omega}\left[\frac{16}{81}J'^2(p_2^0)^2 - \frac{4}{81}K'^2(p_2^2\cos 2l)^2\right]d\omega . \tag{11}$$

The surface integral of any zonal harmonic is

$$\frac{4\pi}{2n+1} ,$$

and for a tesseral harmonic it is

$$\frac{2\pi}{2n+1}\frac{(n-s)!\,(n+s)!}{(n!)^2} ,$$

where n is the degree of the harmonic, and s is the order.[21]

Substituting these values into equation 11 and integrating, we find

$$\overline{(\Delta I)^2} = \frac{16}{405}\left[J'^2 - \frac{3}{4}K'^2\right] ,$$

$$= 0.000\ 000\ 004\ 38 . \tag{12}$$

This estimate was made with only two degrees of freedom; it depends on J' and K'. In effect, we forced the mean value of ΔA, ΔB, and ΔC to be zero. The expected random variations if we do not make this requirement, i.e., with three degrees of freedom, are greater by the ratio $\frac{3}{2}$; hence, the standard deviation of I, as judged from J' and K' is given by

$$\sigma_I^2 = \frac{3}{2}(0.000\ 000\ 004\ 38) ,$$

$$\sigma_I = 0.000081 .$$

To compare this with the theory, we imagine the body of the moon constituted of n lumps of equal volume v.[20] If the lumps were of equal mass m, the moment of inertia I would be given by

$$I = \sum_{i=1}^{n} m\,\rho_i^2 , \tag{13}$$

where ρ_1 is the distance of the ith lump from the axis around which the moment is taken.

Assume, however, that there is a variation Δm in the mass of any individual volume. If the standard deviation of m is σ_m, then it can be shown, by dividing the mass into cylindrical shells of constant ρ_i, that the expected standard deviation of I will be given by

$$\sigma_I^2 = \sum_{i=1}^{n}(\sigma_m\rho_i^2)^2$$

$$\approx \sigma_m^2\sum_{i=1}^{n}\rho_i^4 , \tag{14}$$

where we neglect a factor of $n/(n-1)$ which arises from the fact that the deviations of m are referred to the mean.

To evaluate the sum, which we denote by S, we convert it to a triple integral, noting that

$$v = \frac{V}{n} , \tag{15}$$

where V is the total volume. Then

$$S = \sum\rho_i^4 = \frac{n}{V}\sum_{i=1}^{n}\rho_i^4 v$$

$$\approx \frac{n}{V}\iiint_{V} p^4 dv , \tag{16}$$

over the moon. The triple integral may be obtained from elementary methods by using trigonometric functions (Peirce's tables, Integral No. 273); it is found that

$$S = \frac{8}{35}na^4 . \tag{17}$$

Therefore

$$\sigma_I^2 = \sigma_m^2\left(\frac{8}{35}na^4\right) .$$

The moment of inertia I of a homogeneous sphere around an axis is

$$I = \frac{2}{5}Ma^2 = \frac{2}{5}nma^2 ;$$

therefore

$$\frac{\sigma_I^2}{I^2} = \frac{10}{7}\left(\frac{1}{n}\frac{\sigma_m^2}{m^2}\right) . \qquad (18)$$

The comparable equation

$$a = \pm A\sqrt{n} , \qquad (19)$$

in Urey's notation[20] then requires a correction factor of $(10/7)^{1/2}$. We now apply the same method to the calculation of the x-coordinate of the center of mass, defined as

$$\bar{x} = \frac{m_1x_1 + m_2x_2 + m_3x_3 \cdots}{M} , \qquad (20)$$

where M is the total mass of the moon. Since the square of the standard deviation of $\bar{x}$ is the sum of the squares of the standard deviations of the terms which make it up,

$$\sigma_{\bar{x}}^2 = \left(\frac{\sigma_m x_1}{M}\right)^2 + \left(\frac{\sigma_m x_2}{M}\right)^2 + \left(\frac{\sigma_m x_3}{M}\right)^2 + \cdots ,$$

or

$$\sigma_{\bar{x}}^2 = \frac{\sigma_m^2}{m^2 m^2}\sum_{i=1}^{n} x_i^2 . \qquad (21)$$

As before we convert the summation to an integral over the volume, setting

$$x = a\cos\phi , \qquad (22)$$

where ϕ is colatitude, and using the Peirce tables, Integral Number 273, we obtain

$$\frac{\sigma_{\bar{x}}^2}{a^2} = \frac{1}{5}\left(\frac{1}{n}\frac{\sigma_m^2}{m^2}\right) . \qquad (23)$$

Dividing this equation by equation 18, we find the ratio

$$\frac{\frac{\sigma_{\bar{x}}^2}{a^2}}{\frac{\sigma_I^2}{I^2}} = \frac{7}{50} ; \qquad (24)$$

therefore

$$\frac{\frac{\sigma_{\bar{x}}}{a}}{\frac{\sigma_I}{I}} = 0.374 . \qquad (25)$$

The actual numerical values are:

$$\left.\begin{aligned}\sigma_{\bar{x}} &= 1 \text{ km}\\ a &= 1735\end{aligned}\right\} \quad \frac{\sigma_{\bar{x}}}{a} = 0.00058 ,$$

$$\left.\begin{aligned}\frac{\sigma_I}{Ma^2} &= 0.000081\\ \frac{I}{Ma^2} &= 0.40\end{aligned}\right\} \quad \frac{\sigma_I}{I} = 0.00020 ,$$

$$\frac{\frac{\sigma_{\bar{x}}}{a}}{\frac{\sigma_I}{I}} \doteq 2.9 .$$

The value of the last ratio is 7.8 times larger than the theoretical one (equation 25). It indicates that the gravitational field is smoother than would have been expected on the basis of the displacement of the centers. Whether the difference is significant will be examined below; for the moment, the important point is that the expected ratio is a pure number, dependent neither on the number of lumps assumed nor on their dispersion in mass.

Urey, Elsasser, and Rochester[20] estimate the value of n, the number of separate lumps, from their equivalent of equation 18, in which

$$\frac{\sigma_I}{I} = 0.00020 ;$$

and they have assumed

$$\frac{\sigma_m}{m} = 0.042 .$$

They find $n = 29{,}000$. Substituting in our equation, we find 63,000 separate lumps. With either Urey's or our estimate of the effect of Urey's hypothesis the probability is of the order of 10^{-5} that the same set of lumps could give a displacement of the centers which is over seven times the theoretical value.

If we calculate n from equation 23, we get

about 1100 lumps, corresponding to a sevenfold larger value of σ_I/I. The probability of getting an observed value of σ_I/I which is $\frac{1}{7}$ of the expected one is about 0.1; it is thus far greater than that of getting a deviation 7 times the expected one. On the other hand, there are *two* determinations of the deviation of the moments of inertia from regularity, so that the overall probability is about 0.01.

Obviously the most probable situation is that the true number n is between 1100 and 63,000, but it must be nearer to 1100 than to 63,000. A situation of this kind is conveniently handled by the F-ratio test. To apply this F-ratio test, the ratio F of the two variances (i.e., the squares of the dispersions) is formed; and the probability that these two variances are significantly different is taken from a triple entry table, whose entries are F, n_1 (the number of degrees of freedom in the numerator), and n_2 (the number of degrees of freedom in the denominator).

The term "number of degrees of freedom" is a number of different measurements which have made on the quantity whose dispersion is to be determined. The number of degrees of freedom is one less than the number of measurements made; clearly, if we have only one measurement of a given quantity to work with, we have no idea of the dispersion; if we have two measurements, we have only one difference to work with, and so on. Here we are estimating the quantity

$$\frac{1}{n}\frac{\sigma_m^2}{m^2} ,$$

in the first place, on the basis of the difference in position between the center of figure and the center of gravity. This amounts to two measurements or one difference, and hence one degree of freedom. We have

$$\frac{1}{n}\frac{\sigma_m^2}{m^2} = \frac{7}{10}\left(\frac{\sigma_I}{I}\right)^2 = 0.000\ 000\ 028 .$$

In the second place, we are estimating the same quantity on the basis of two differences, J' and K', among three quantities, A', B', and C'; hence we are estimating with two degrees of freedom. We find

$$\frac{1}{n}\left(\frac{\sigma_m}{m}\right)^2 = 5\left(\frac{\sigma_E}{a}\right)^2 = 0.000\ 001\ 68 .$$

The ratio F of the two determinations of σ_m^2 is therefore 58.0. Considering the F test table, with one degree of freedom for the greater variance, and two degrees of freedom for the smaller, we find that the probability of such an occurrence by chance is between 1 and 5 percent.

It is thus clear that the displacement between the center of the moon's figure and its center of gravity does actually suggest the existence of isostasy.

The differences of the moments of inertia imply stress differences deep in the moon which are of the same order of magnitude (about 2×10^7 dynes/cm^2) as those implied by the nonhydrostatic components of the earth's gravitational field. These were worked out by Jeffreys[21] who started with gravimetric determinations of the harmonics; but his estimates of the general size of the harmonics are so close to those found from satellite observations that his conclusions can be applied without change.

Some minor features of the moon's surface likewise suggest isostatic readjustment. In particular, many maria are surrounded by evidence of tectonic readjustment along their shores; specifically:

1. More or less circular systems of rilles surround Mare Crisium, Mare Imbrium, Mare Serenitatis, and Mare Tranquillitatis;
2. Fault scarps are visible on the shores of Mare Humorum and Mare Nubium, the downthrow sides being toward the maria;
3. Most of the maria have craters along their shores which dip downward toward them. In the case of Mare Humorum, two of these craters have actually been sliced in two by a fault; the same is suspected of the Straight Wall.

The evidence of tectonic instability and, especially, of sinking, means that the maria are more than just holes blasted by impacts in an essentially homogeneous moon. For in the latter case, hydrostatic forces would be pushing the maria upwards to seek the same level as the continents. The fact that, although lower than the continents, they tend to sink, suggests that they are underlain by heavier material than are the continents.

A possible alternative reason for sinking might be that the maria are enormous calderas of collapse; if the dark material of the maria has been

withdrawn from the interior in any manner, certainly collapse would be expected over large areas. This cannot, however, be the whole story of the origin of the maria, since in this case the material withdrawn would be sufficient to fill them up. Moreover, the liquefaction, or the rendering plastic, of so large a body of rock probably implies a degree of weakness, in the crust as a whole, which is hard to reconcile with the mechanical strength necessary to support the difference of level between the continents and maria.

Laccoliths

The second manifestation of granitic rock which ought to produce observable effects on the moon is the production of laccoliths. Laccoliths are moderate-sized intrusions; on the earth they are usually found in sedimentary rocks, where a weaker layer permits the magma to spread laterally more easily than by forcing its way vertically upward. Laccoliths tend to produce surface bulges which are generally round in plan, with a diameter of the order of 10 km, and gentle external slopes. Basaltic magmas do not ordinarily produce laccoliths because of their low viscosity; instead of accumulating around the pipe through which they came, the basaltic magmas usually spread laterally forming the sills. Thus the presence of laccoliths indicates viscous, and hence silicic, magmas.

On the moon the structures called domes by Spurr[22] and others correspond in many respects to terrestrial laccoliths. They are normally found in the maria, though observational selection may play a part here, since it is hard to detect a gentle gradation in a rough area; their sizes, up to a diameter of 10 km, are similar to those of terrestrial laccoliths; the slopes are also similar. About $\frac{2}{3}$ of the domes have summit craters, as do certain terrestrial laccoliths described by Daly.[23]

Many authorities, including Salisbury, Spurr, Kuiper, Shoemaker and Hackman, and Gold, do not consider the domes to be laccoliths.[22,24–27] A principal objection is the requirement for a layered rock, which on earth usually means a sedimentary rock. If the maria consist of basaltic flows, then weak strata are not to be expected. If they are composed, as Gold has suggested,[27] of dust derived from high ground, then they ought not to contain especially strong layers. If, however, they consist of ash flows laid down episodically, then strata of varying strength may be expected because of the welding of the central portions of ash flows. This point will be considered further in the discussion of ash flows. It may be concluded that lunar evidence is consistent with the presence of laccoliths, but does not unambiguously indicate that they are present.

Surface flows of acid rocks tend to have irregular upper surfaces because of their higher viscosity.[28] A possible example of a rough surface on a recent flow is west of Copernicus (about 8°E, 4°N). This dark area seems to interrupt some of the rays from Copernicus; it may therefore be one of the very latest events in the moon's evolution.

Peleean Volcanism

At the surface the principal effects of the extrusion of granitic rock may be classified under the head of Peleean volcanism, named after Mont Pelée in Martinique. (The Hawaiian fire goddess Pele has also had at least one phenomenon, Pele's hair, named after her; this is confusing because the Hawaiian volcanoes are typical of the basaltic group.) Peleean volcanism[29] is characterized by at least three phenomena which might be distinguished on the surface of the moon. These are:

1. The extrusion of lava in the form of ridges, spines, and tholoids (small, steep-sided domes). These appear to consist of very viscous lavas extruded under pressure, through cracks and holes in the crust. The most remarkable case is that of a spine extruded to a height of 300 meters from the crater of Mont Pelée within a period of a few months.

2. The production of ash: this is divided into:

a. Ash falls, like the famous eruption of A.D. 79 which covered Pompeii.

b. The remarkable—but by no means uncommon—phenomenon of ash flows.

These effects will be considered in order. Under the extrusion of lava would be included some small swellings very common on the moon, a kilometer or two in diameter, which may be compared with the terrestrial spines and tholoids. On the moon, such swellings are common in the area surrounding Mare Imbrium; Urey considers them rubble

thrown out of the crater at the time of its formation by collision.[19] If they are rubble, however, it is difficult to see why this material failed to produce the characteristic elongated craters which are found in such numbers associated with Tycho and Copernicus. In these latter, relatively certain cases, the relief produced is negative; it is hard to see why the Imbrian ejecta should produce positive relief, especially when the pieces had farther to go and hence possessed more energy per gram.

Shaler compared the ridges which are a common feature of the central belt of the moon, south of Mare Imbrium, with the ridges formed on the earth by the extrusion of viscous (he mentioned trachytic) lavas.[30] These structures are most clearly seen on the moon near Boscovich and Julius Caesar, and in the Haemus mountains. They run parallel to a system of narrow valleys found in the same areas, which diverge from the center of Mare Imbrium. These features have been attributed, by Gilbert[31] and Urey,[19] to scoring by masses fired from Mare Imbrium. Once again the parallel with large craters is missing. The long groove-like valleys are entirely unlike the short elliptical craters, for example, those associated with Copernicus.

The central peaks of certain craters constitute a special group of hills associated with the aforementioned ridges. These peaks were regarded as extrusions by Shaler[30] and Kuiper.[25] This view does not necessarily conflict with impact theories of the origin of the craters; on the earth such tholoids are common along cracks of an entirely independent origin.[29] It is interesting to note that the central peak of Alphonsus, like several other peaks, is aligned toward the center of Mare Imbrium, although it is mechanically impossible that it could have been shaped by fragments from the Imbrian center. The tectonic significance of the direction toward the center of Mare Imbrium is further emphasized by the Straight Wall, which is a clearly defined fault not far from Alphonsus. It is, in fact, clear that such ridges are manifestations of one of the lunar tectonic grids on which Fielder has laid emphasis.[32] This grid is clearly related to Mare Imbrium in a subtler way than by mechanical scoring.

With reference to the ejection of volcanic ash, there is a little evidence for ash *falls* on the moon. A few craters, notably several inside Alphonsus, are surrounded by darkened areas which may represent ash falls. One small crater, approximately 2 diameters of the crater Mairan east of it, appears to be a typical ash cone, with slopes of about 38 degrees (E. A. Whitaker, verbal communication).

Terrestrial ash *flows* (which are quite different from ash *falls*) consist of avalanches of volcanic ash which are fluidized by contained gases.[33] Several features of an ash flow are of especial importance:

1. The total emplacement, which may be comprised of one or more individual ash flows, is very extensive, so that on the earth the area covered by ash flows may be fully as great as that covered by basaltic flows.* Individual units cover tens of thousands of square miles;[34]
2. The surface is level and remarkably smooth;[35]
3. The edges, unlike the edges of basaltic flows, feather out against the surrounding topography;
4. The top surface is composed of loose ash;[35]
5. The ash bed is welded into a solid obsidian in the middle, where the heat has stayed longest in the ash—not at the bottom where it escapes into the ground, nor at the top where it escapes into the air. The welding is such that it converts loose ash into a rock of unusual strength and toughness;
6. The collapse of the porous structure is greatest where the ash is deepest. Hence the surface of an ash flow tends to show the same features as the underlying original surface, although much diminished in height and depth;[28]
7. The chemical behavior of the interior depends in part on the gases which pass up through it; hence the chemistry at any point depends on the depth (verbal communication from Michael Bickerman, student at Arizona State University).

These features strengthen the suggestion tenta-

*It is only rather recently that geologists in general have recognized the extent and importance on earth of ash flows that produce welded tuffs (ignimbrite). According to Ross and Smith (reference 33), "Many ash flow tuffs have not been mapped as such in the past because they were not recognized and many areas previously mapped as lava flows are now known to be ash flow tuffs. Other areas will doubtless be found after more detailed laboratory study and areal mapping."

tively put forward by Shoemaker and Hackman[26] that the maria might possibly be regarded as ash flows (ignimbrite). This idea is very close to Gold's hypothesis that the maria are fluidized dust.[27] In support of his idea, Gold pointed out radar evidence that the surface is quite level on the scale of the radar wavelength, i.e., around 10 cm. This evidence has been further supported by the studies of Senior and Siegel on the significance of polarization in radar returns.[36] It disagrees entirely with the expected surface of a basaltic flow, which is likely to be blocky on a scale of a meter or so. Where comparatively smooth surfaces are found on lava flows, they are the result of aqueous erosion which has either removed the blocky crust or filled the cracks with silt.

Gold further pointed out the significance of the absence of scarps at the edges of the maria, and of the thermal indications of a loosely compacted surface.

On the other hand, the welding of the ash flows in horizontal strata is not predicted by Gold's mechanism of dust fluidization; and an effect such as this is necessary if we are to understand the domes as laccoliths. The alternation of loose ash and welded tuff which would be expected from the ash flow hypothesis does provide the necessary stratification; but if the dust moves continuously into the maria, then it is hard to see why there should be the pronounced stratification which is required.

The tendency of ash flows to reproduce the underlying topography on their surfaces recalls the phenomenon of ghost craters. On certain maria, it is possible to trace the outlines of craters that apparently are mostly below the surface. Sometimes a crater can be traced from the mare border inward, so that we can be sure that what we see on the mare is really the outline of a crater. The crater is revealed both by its effect on the surface elevations and by its effect on the coloration of the surface. The latter is apparently due to the effect pointed out by Bickerman—the dependence of the chemical state of the tuff on its depth.

Since ash flows are supported by gas, it is necessary to ask whether such eruptions could travel very far in the near-vacuum at the moon's surface. It might seem at first that the flow would collapse rapidly because of the escape of gas into space. This possibility is strongly implied by McTaggart,[37] who suggested that the mobility of ash flows is due primarily to air entrapped by the advancing front. If McTaggart is correct, we cannot appeal to ash flows as a means of covering large areas of the moon's surface; we shall therefore discuss his treatment in some detail.

McTaggart's analysis considers a large block supported by gases flowing from beneath; it is based on Torricelli's formula for the laminar flow of an inviscid fluid through an orifice. The inertial terms in the flow equations are taken into account. This analysis is believed to rest on a misapprehension of the problem. The fundamental differential equation for flow in the x direction, at the velocity u, is

$$\frac{\partial p}{\partial x} = \rho \frac{du}{dt} + \rho X + \mu \nabla^2 u , \qquad (26)$$

where p is the pressure, ρ the density, du/dt the time rate of change of velocity (since we are following the flow, i.e., for a chosen element moving with the fluid), X the body force in the x direction per unit mass, and μ the viscosity. McTaggart has neglected the second and third terms on the right-hand side, and has, by his use of Torricelli's equation, in effect equated the left-hand side to the first term on the right. Physically, this means that he assumed the particles are suspended by the inertial reaction of the gases as they speed up in passing through the passages around the particles. McTaggart correctly shows that this force is entirely inadequate to support the materials.

If, however, the materials are finely divided, then a different and more effective mechanism comes into play, namely the viscous resistance of the air. Since the passages between the particles are narrow and the layer next to the particles is at rest with respect to them, there are steep velocity gradients and the viscosity effects are large. Numerical estimates show that in this case it is correct to regard the first two terms on the right-hand side of this equation as negligible, by comparison with the third, and write

$$\frac{\partial p}{\partial x} = \mu^2 u . \qquad (27)$$

From this equation, Lamb derived the Poiseuille

law for the flux ϕ through a capillary tube (reference 38, p. 585, equation 4):

$$\phi = \frac{\pi a^4}{8\mu}\frac{\Delta p}{L}, \tag{28}$$

where a is the radius and L the length. If we consider instead the average velocity u^0, so that

$$\phi = \pi a^2 u_0, \tag{29}$$

then

$$\Delta p = \frac{32\mu u_0 L}{D_t^2}, \tag{30}$$

where D_t is the tube diameter. It is in this form that Leva quotes the Poiseuille equation (reference 39, p. 44).

The weight of the column of material which the pressure must support is

$$L(\rho_s - \rho_f)(1 - \epsilon)g = \Delta p, \tag{31}$$

where g is the acceleration due to gravity. Here the quantity $(\rho_s - \rho_f)$ represents the density of the solid particles minus the buoyancy effect (negligible for gases); and ϵ represents the fraction of a given volume of the aggregate which is occupied by voids, so that $\rho_s(1-\epsilon)$ is the average density of the aggregate. Equating the two expressions for Δp and replacing $\rho_f u_0$ by the rate of flow G in grams per second per cm^2, we have

$$G = \frac{1 - \epsilon}{32}\frac{D_t^2 \rho_f (\rho_s - \rho_f) g}{\mu}. \tag{32}$$

In practice we can only hope to find a law of the above *form*. We are not given the tube diameter D_t, but only the related particle diameter D_p. The passages are crooked, and thus the length L in equation 30 is not really the same as in equation 28. The effect of particle roughness will be felt both directly and through its effect on the voidage ϵ.

Leva has demonstrated the remarkable and fortunate fact that over a wide range of densities, flow rates, particle sizes, and fluid viscosities of both gases and liquids, the minimum flow G_{mf} required to fluidize an aggregate is

$$G_{mf} = C\frac{D_p^2 \rho_f (\rho_s - \rho_f) g}{\mu}, \tag{33}$$

where C is not quite constant, but a slowly varying function of the Reynolds number R:

$$C \propto R^{-0.063},$$

where

$$R = \frac{GD_p}{\mu} \tag{34}$$

as usual. The resulting empirical equation in cgs

TABLE 1.—*Values of parameters for application of equation 35*

Parameter	Terrestrial Case	Lunar Case
Bed thickness	100 m	100 m
Temperature	850°C	850°C
Percentage water in rock	0.1 percent	0.1 percent
Particle diameter	0.01 cm	0.01 cm
Gravity	980 gal	162 gal
Gas (steam) density	1.7×10^{-2} gm/cm^3	2.5×10^{-3} gm/cm^3
Viscosity	4.2×10^{-4} po	4.2×10^{-4} po
Solid density	2.5 gm/cm^3	2.5 gm/cm^3
Minimum gas flow	7.6×10^{-3} gm/cm^2-sec	2.3×10^{-4} gm/cm^2-sec
Duration of flow	0.8 hr	27 hr
Extent of flow	80 km	2700 km

units, if we neglect buoyancy, is

$$G_{mf} = 1.09 \times 10^{-3} \frac{D_p^{1.82} (\rho_f \rho_s g)^{0.94}}{\mu^{0.88}} . \quad (35)$$

We will now apply this equation to two cases, a terrestrial and a lunar case. The results are summarized in table 1. The assumed bed thickness is 100 meters, in agreement with typical terrestrial ash flows. The temperature of 850°C is that determined by Boyd.[28] The percentage of water in the rock (0.1 percent) is near the minimum for terrestrial granites. Most granites have values nearer 1 percent; tektites, however, probably owing to the heating to which they have been subjected, often have as little as 0.002 percent, though Friedman[15] has measured values as high as 0.014 percent in philippinites.

The particle diameter is toward the lower end of the size distribution observed. The reason is that the size of the passages between the particles is controlled, not by the large particles, but by the small ones. The pressure, not listed in table 1, is calculated for the midpoint of the bed from equation 31, by assuming a atmosphere pressure at the surface in the terrestrial case, and 0 atmosphere in the lunar case. From the pressure and temperature, the density and viscosity are found with the usual tables. The solid density corresponds to an acid rock.

The minimum gas flow required for fluidization can be calculated from Leva's equation. The duration of flow is estimated by comparing the rate of flow for minimum fluidization with the available amount of water. The extent of flow is calculated on the assumption that it moves at the rate observed on earth, about 100 km/hr.

The greater duration and extent of the lunar flows are the result of the lower gravity, which acts in two ways: (1) Directly, by reducing the weight of the layers to be supported; (2) Indirectly, by reducing the pressure at every depth, and hence the density, and thus the mass of fluid required. (The supporting power of a gas flow depends on its volume and its viscosity; since the viscosity is independent of the density, a flow of low density gives the same support as one of high density and is more economical in regard to material.) The combination of these two effects produces a factor of about 32 in favor of the greater duration and extent of lunar ash flows.

The absence of an atmosphere on the moon is almost irrelevant, since all the equations for the support of the layer involve only a pressure difference. However, in calculating density at a given temperature, the absolute value of the pressure must be used; and here it is clear that the lunar flows will require less gas than terrestrial flows, since they will have lower pressures and densities near the surface. Surprisingly, lunar conditions are actually more favorable to deposition of extensive tuffs from ash flows than are terrestrial conditions.

CONCLUSION

It appears that it is quite possible to furnish explanations of the observed lunar land forms on the assumption that the moon's surface has a high proportion of granitic rocks. The explanations are not always the accepted ones; but they are well within the boundaries of physical plausibility. They do not appear to conflict with Dollfus' polarization observations of the moon, or with his conclusions as to the volcanic nature of the lunar surface.[40]

ACKNOWLEDGMENTS

The authors are grateful to Dr. H. C. Urey, who provided the stimulus for this paper through his publications and private discussions. We also express our appreciation to Dr. D. W. G. Arthur, Dr. F. R. Boyd, Dr. P. Damon, Dr. J. Green, Mr. R. J. Hackman, Dr. E. M. Shoemaker, Mr. R. L. Smith, and Mr. E. A. Whitaker for valuable discussions on various aspects of this paper. And we especially thank our colleague Mr. P. D. Lowman, Jr. for his many invaluable discussions and criticisms of the geological content of this report.

REFERENCES

1. Ninninger, H. H., "The Moon as a Source of Tektites, Pt. 1," *Sky and Tel.*, **2**(4): 12–15, February 1943; "The Moon as a Source of Tektites, Pt. 2," *ibid.*, **2**(5): 8–9, March 1943.
2. Dietz, R. S., "The Meteoritic Impact Origin of the Moon's Surface Features," *J. Geol.*, **54**(6): 359–375, November 1946.

3. O'KEEFE, J. A., "Origin of Tektites," *Nature*, **181** (4603): 172–173, January 18, 1958; see also "The Origin of Tektites," NASA Technical Note D-490, November 1960.
4. VARSAVSKY, C. M., "Dynamical Limits on a Lunar Origin for Tektites," *Geochim. et Cosmochim. Acta*, **14**(4): 291–303, 1958.
5. GOLD, T., "Origin of Tektites," *Nature*, **181**(4603): 173–174, January 18, 1958.
6. CHAPMAN, D. R., "Recent Re-Entry Research and the Cosmic Origin of Tektites," *Nature*, **188**(4748): 353–355, October 29, 1960.
7. VERBEEK, R. D. M., "Over Glaskogels van Billiton," *Verslagen van de Gewone Vergaderingen der Wis- en Natuurkundige Afdeeling, Koninklijke Akademie van Wetenschappen te Amsterdam*, **5**, 421–425, 1897.
8. GILVARRY, J. J., "Origin and Nature of Lunar Surface Features," *Nature*, **188**(4754): 886–891, December 10, 1960.
9. LOWMAN, P. D., Jr., "Tektites vs. Terrestrial Rocks: A Comparison of Variance in Compositions," *Geochim. et Cosmochim. Acta*, **26**, 561–579, May 1962.
10. TUTTLE, O. F., and BOWEN, N. L., "Origin of Granite in the Light of Experimental Studies in the System $NaAlSi_3O_8$–$KAlSi_3O_8$–SiO_2–H_2O," New York: Geological Society of America, 1958 (Geological Society of America, Memoir 74).
11. BARNES, V. E., "North American Tektites," in: *Contributions to Geology*, 1939, Part 2, The University of Texas Press, **3945**, 477–582, 1939.
12. LOWMAN, P. D., Jr., "The Relation of Tektites to Lunar Igneous Activity," *J. Geophys. Res.*, **67**(4): 1646, April 1962 (Abstract).
13. COHEN, A. J., "Trace Element Relationships and Terrestrial Origin of Tektites," *Nature*, **188**(4751): 653–654, November 19, 1960.
14. TAYLOR, S. R., "Distillation of Alkali Elements During Formation of Australite Flanges," *Nature*, **189**(4765): 630–633, February 25, 1961.
15. FRIEDMAN, I., "The Water, Deuterium, Gas and Uranium Content of Tektites," *Geochim. et Cosmochim. Acta*, **14**(4): 316–322, 1958.
16. SENFTLE, F. E., and THORPE, A., "Magnetic Susceptibility of Tektites and Some Other Glasses," *Geochim. et Cosmochim. Acta*, **17**(3/4): 234–247, 1959.
17. BALDWIN, R. B., "A Lunar Contour Map," *Sky and Tel.*, **21**(2): 84–85, February 1961.
18. SCHRUTKA-RECHTENSTAMM, G., "Neureduktion der 150 Mondpunkte der Breslauer Messungen von J. Franz," *Sitzungsberichten der Österreichischen Akademie der Wissenschaften, Mathematisch-Naturwissenschaftliche Klasse, Abteilung II*, **167**(1–4): 71–123, 1958.
19. UREY, H. C., "The Planets: Their Origin and Development," New Haven: Yale University Press, 1952.
20. UREY, H. C., ELSASSER, W. M., and ROCHESTER, M. G., "Note on the Internal Structure of the Moon," *Astrophys. J.*, **129**(3): 842–848, May 1959.
21. JEFFREYS, H., "The Earth: Its Origin, History, and Physical Constitution," 4th Ed., Cambridge University Press, 1959, Chapter IV.
22. SPURR, J. E., "Geology Applied to Selenology," Lancaster, Pennsylvania: Science Press Printing Co., 1944, p. 83.
23. DALY, R. A., "Igneous Rocks and the Depths of the Earth," New York: McGraw-Hill, 1933, p. 153.
24. SALISBURY, J., "Origin of Lunar Domes," *Proc. Lunar and Planet. Exploration Colloq.*, **2**(2): 22–26, March 17, 1960.
25. KUIPER, G. P., "The Exploration of the Moon," in: *Vistas in Astronautics*, ed. by M. Alperin and H. F. Gregory, New York: Pergamon Press, 1959, Vol. 2, pp. 273–313.
26. SHOEMAKER, E. M., and HACKMAN, R. J., "Stratigraphic Basis for a Lunar Time Scale," in: *Astrogeologic Studies, Semiannual Progress Rept., August* 25, 1960 *to February* 25, 1961, U.S. Geological Survey, March 1961, p. 21.
27. GOLD, T., "The Lunar Surface," *Monthly Not. Roy. Astronom. Soc.*, **115**(6): 585–604, 1955.
28. BOYD, F. R., "Welded Tuffs and Flows in the Rhyolite Plateau of Yellowstone Park, Wyoming," *Bull. Geol. Soc. Amer.*, **72**(3): 387–426, March 1961.
29. COTTON, C. A., "Volcanoes as Landscape Forms," 2nd Ed., Christchurch: Whitcombe and Tombs, 1952.
30. SHALER, N. S., "A Comparison of the Features of the Earth and the Moon," Washington: Smithsonian Institution, 1903 (*Smithsnoian Contributions to Knowledge*, Part I, Vol. 34).
31. GILBERT, G. K., "The Moon's Face; A Study of the Origin of Its Features," *Bull. Phil. Soc. Washington*, **12**, 241–292, 1892–1894.
32. FIELDER, G., "Structure of the Moon's Surface," New York: Pergamon Press, 1961.
33. ROSS, C. S., and SMITH, R. L., "Ash-Flow Tuffs: Their Origin, Geologic Relations, and Identification," Washington: U.S. Government Printing Office, 1961 (U.S. Geological Survey, Professional Paper 366).
34. SMITH, R. L., "Ash Flows," *Bull. Geol. Soc. Amer.*, **71**(6): 795–841, June 1960.
35. GRIGGS, R. F., "Our Greatest National Monument," *Nat. Geog. Mag.*, **40**(3): 219–292, September 1921.
36. SENIOR, T. B. A., and SIEGEL, K. M., "Radar Reflection Characteristics of the Moon," in: *Paris Symposium on Radio Astronomy*, ed. by R. N. Bracewell, Stanford: Stanford University Press, 1959, pp. 29–46.
37. MCTAGGART, K. C., "The Mobility of Nuées Ardentes," *Amer. J. Sci.*, **258**(5): 369 382, May 1960.
38. LAMB, H., "Hydrodynamics," 6th Ed., Cambridge University Press, 1932; New York: Dover, 1945, p. 585, Equation 4.
39. LEVA, M., "Fluidization," New York: McGraw-Hill, 1959.
40. DOLLFUS, A., "La Courbe de Polarisation de la Terre et la Nature du Sol Lunaire," *Comptes Rendus, Académie des Sciences* (Paris), **235**(18): 1013–1016, November 3, 1952.

ORIGIN OF TEKTITES

JOHN A. O'KEEFE

AND

BARBARA E. SHUTE

Goddard Space Flight Center

A comet of the size recently postulated by H. C. Urey would leave a large crater. It is shown, from aerodynamic theory, from observations of distribution around terrestrial impact craters, and from experimental nuclear explosions, that the observed distribution of tektites cannot be the result of impact on the earth, whether cometary or meteoritic. It is further shown, from aerodynamic theory, from observation of a meteor shower, and from study of the breakup of artificial satellites, that the distribution of tektites can be accounted for as a result of fusion stripping of a satellite, as originally suggested by Suess.

Urey[1] has recently rediscussed the problem of the origin of tektites in the light of new evidence. He shows that it is not reasonable to think of tektites as formed individually by impact at the moon's surface, since in this case the tektites would undoubtedly be scattered more or less uniformly over the surface of the earth, and through at least the Cenozoic strata, which is not observed We agree with this argument, and we further agree with his opinion that the whole Far Eastern strewnfield, from China to Tasmania, is to be regarded as a single event.

Unfortunately, it appears that his hypothesis of the origin of tektites by cometary impact on the earth contains contradictory elements. On the one hand, it is asserted that the atmosphere arrests the cometary head as it descends, so that the primary effects are not a shock-produced crater in the solid ground, but a mass of heated gas. On the other hand, it is supposed that the tektites produced on the ground by this heated air are not arrested, but rise to the top of the atmosphere with ballistic velocity sufficient to carry them thousands of kilometers.

The laws of aerodynamics do not work this way. It is the small bodies which are stopped by the atmosphere, and the big bodies which get through. The drag pressure is given by

$$p = C_d \tfrac{1}{2} \rho V^2$$

where ρ is the density of the air, V the velocity of the body relative to the air, and C_d the drag coefficient. The drag coefficient is of the order of 1, and will be omitted from the rest of the discussion, since we are aiming at the order of magnitude. If the area of the body is A, and the increment of distance traversed is ds, then the increment of work, dW, is

$$dW = pA\ ds = \tfrac{1}{2} \rho A V^2\ ds$$

When the work done becomes of the order of magnitude of the initial kinetic energy, $\frac{1}{2}MV^2$, (M being the mass of the body), then the body is essentially stopped. Neglecting the variation in velocity, this means

$$\tfrac{1}{2} M V^2 = \int_{S_1}^{S_2} dW = \tfrac{1}{2} \int_{S_1}^{S_2} \rho A V^2\ ds$$

that is,

$$M = A \int_{S_1}^{S_2} \rho\ ds$$

that is, when the mass of the air encountered is

equal to the mass of the body. This principle, though not this derivation, was stated to us by F. L. Whipple.

Since a vertical atmospheric column has about 1 kg of mass per square centimeter, it is to be expected that bodies with less than this mass per square centimeter of frontal area will be arrested. In practice, this means that bodies with a diameter less than something like 5 m will be stopped by the atmosphere, and will reach the ground with terminal velocity. Larger bodies will penetrate and will make craters. This expectation is approximately satisfied by the facts about the largest meteorites and the smallest craters.

The general principle at work here can also be derived from Newton's Third Law of the conservation of momentum. Alternatively, from purely dimensional considerations, it is clear that the total drag must increase with the square of the linear dimensions, while the mass, and hence the energy, increase with the cube; hence once more we see that the larger bodies must be the ones which will penetrate, while the smaller bodies will be stopped.

Even if the density of the cometary head is as little as 0.01 g/cm^3, and the diameter is 10 km, as Urey[2] has previously suggested, the mass per square centimeter of frontal area will be much greater than that of the atmosphere, and the body will be stopped, not by the atmosphere, but by the earth. The comet postulated by Urey would have an energy of 5×10^{28} ergs.

On the other hand, Shoemaker finds[3] that an energy only a little more than the above, namely, 7.5×10^{28} ergs, was required for the formation of the lunar crater Copernicus, 80 km in diameter, with walls 4 km above the floor. Hence we would expect that a conspicuous terrestrial crater would have been formed by the impact which, on Urey's theory, produced the Far Eastern strewnfield. The crater would presumably be nearer the northern end of the field, since the tektites are much more numerous there. It would be marked by a large circular lake. No such lake can be found, however, either in Laos, or in Thailand, or Burma, or Yunnan Province, China. It happens that Yunnan Province is covered by a 1:50,000 map series which one of us personally examined during World War II and compared with Army Air Force astronomical positions. The series is adequate to show a lake of this size, which would, in fact, cover a dozen sheets of the map. The lake is not there.

Urey also considered a comet 70 kilometers in diameter, which would produce a lake 7 times wider. This is excluded a fortiori.

In any case, why a comet? There are nickel-iron spherules in tektites, but no volatiles; hence, one's first guess would be a meteorite. The spherules make it reasonably sure that the impacting body did mix physically with the ground which it struck; then why the mechanism of compressed hot gases to keep the two apart?

Consider next the second postulate of the cometary theory, which it shares with all theories of terrestrial origin, namely, that tekitites were melted by impact and then ejected through the atmosphere. The very small mass per square centimeter of frontal area, which never exceeds about 25 g, is far from meeting aerodynamic requirements. Adams and Huffaker[4] have pointed out that a typical tektite would suffer an acceleration of 96,000 times terrestrial gravity. It would be arrested in a very short distance.

To meet this problem Urey suggests that a portion of the atmosphere may be blown outward by the explosion, carrying the tektites with it. Calculation based on the theory of Taylor[5] shows that such a thing can only happen for an explosion of about the size postulated in Urey's first paper, namely about 5×10^{28} ergs. It is necessary that the blast wave carry the tektite with ballistic velocity up to a level of 70 km or so, where further atmospheric resistance can be neglected. Ballistic velocity is here taken as 4.5 km sec^{-1}, the velocity required to span the radius of the Far Eastern strewnfield. It turns out that weaker explosions would decay before reaching the effective limit of the atmosphere.

We have already seen that an impact of this kind is to big to escape notice if produced in the relatively recent past.

The external form of the tektites presents a serious difficulty for all theories which imply that tektites were melted by the impact and then ejected through the atmosphere. They appear to have been large liquid drops whose form is

due to surface tension. In rare cases, they have contained large bubbles within them. Liquid drops of this size are extremely delicate—far more delicate than an egg, for instance—and to find them emerging intact and at ballistic velocity from a great impact, in which rocks are reduced to a fine breccia, would be paradoxical.

In actual fact, the impure glass from the Ries Kessel is found no more than 10 km from the rim. Larger blocks are found at distances up to 70 km. The ejecta at Wabar, Henbury, and the Arizona craters are likewise within 10 km of the crater rim.

On the experimental side, Glasstone[6] gives data on velocities and distribution of particles from atomic explosions. It turns out that particles over 300 μ in diameter are distributed within a very limited radius of the impact. The wide distribution of smaller particles is a consequence of air currents and not of ballistic trajectories. Pebbles and flying solid objects in atomic explosions are mostly the result of the air blast on bodies in the immediate vicinity. A limit on the order of 1 or 2 km appears to be reasonable for material thrown out from the center of even the greatest atomic explosions.

Thus, there is neither theoretical nor observational nor experimental evidence that tektites can be distributed in the observed manner from any reasonable ground impact.

We have suggested that the impact took place on the moon, and that among the ejecta were large solid blocks. Urey inquires whether large blocks could be impelled at velocities of 2.4 km sec^{-1} by impact. Let us note the secondary craters around Copernicus[3] which are apparently produced by much larger blocks than these we have supposed, moving, it is true, with somewhat lower velocities, on the order of $\frac{1}{2}$ to 1 km sec^{-1}.

Urey has suggested that the probability of arrival at the earth in a grazing orbit from the moon is very small. If we consider those bodies which leave the moon with velocities greater than 2.3 km sec^{-1}, and which therefore do not fall back at once, they can be divided into two classes, depending on whether they escape at once from the earth-moon system, or are temporarily trapped. Theoretically there is a third class, which is permanently trapped, at least in the realm of validity of the restricted problem of three bodies; but Kopal has shown[7] that this class corresponds to a negligibly small range of velocities.

Let us define the residual energy as the energy of the body, after subtraction of the energy of escape; and let us further define the residual velocity as the velocity corresponding to the residual energy. The vector sum of the residual velocity and the moon's orbital velocity is the geocentric velocity; if this exceeds 1.4 km sec^{-1} in absolute value, the body will escape the earth-moon system at once, and will go into orbit around the sun. After a very long time it will probably strike the earth. The encounter is not likely to be at grazing incidence, and hence such bodies will not form tektites.

If the absolute value of the geocentric velocity is less than 1.4 km sec^{-1}, then the body will temporarily describe an eccentric orbit around the earth, more or less perturbed by the sun and the moon. A few numerical integrations of this problem are now available; they indicate that under some circumstances bodies in such orbits will be perturbed by the moon in such a way as to produce large, long-term oscillations in the eccentricity. The consequent variation in perigee height is to large, in some cases which have been studied, as to bring the perigee below the surface of the earth. Because of the gradual nature of these perturbations, it is clear that the likelihood of encounter with the atmosphere is much greater than would be expected from purely geometrical considerations. Once the atmosphere is touched, the eccentricity and the semimajor axis of the orbit will be rapidly reduced, the perigee height remaining approximately the same, so that the further influence of lunisolar perturbations can be disregarded. The body will eventually enter the atmosphere along a grazing orbit.

From general considerations, as well as the few numerical integrations now available, it appears that in such orbits, the quantity

$$J_r = (1-e^2)^{1/2} \cos i$$

where i is the inclination of the orbit to the plane of the moon's orbit, will be nearly constant up to the time when the atmosphere is en-

countered. In order to reach sufficiently large values of the eccentrivity, and hence sufficiently small values of $(1-e^2)^{1/2}$ it is necessary that J_r be small; and this is certain to occur if the inclination is large. Numerically, it appears that an inclination to the plane of the moon's orbit greater than 60° is sufficient.

Further work along this line is needed and is being done; but these results show that the situation is not a simple one, and that Urey's geometric approximation is not adequate to give even a rough estimate.

Adams and Huffaker[4] have worked out the mechanics of Suess's suggestion[8] that tektites were formed in the skipping entry of a large parent body in the earth's atmosphere. They find that fusion stripping will work, provided that allowance is made for the heating of the parent body by radiation from the very strong shock which is produced. The parent body must be reasonably transparent to the radiation. Convective heating will not penetrate sufficiently deep.

As noted in Urey's paper, we have managed to show how bodies which are breaking up in orbit around the earth may be distributed over a broad area on the ground. The mechanism which we have proposed is independent of the manner of breakup, provided that it is within the atmosphere. At the present time, we feel that the stage of melting and dripping discussed by Adams and Huffaker[4] may have followed an earlier stage of mechanical breakup, and that the dripping may account both for the indomalaysianites and for the australites as successive stages in the dripping of a single group of bodies. The geometrical situation is not, however greatly altered by this fact. The significant point is that the fragments of a body which breaks up while moving in a moderately elliptical orbit will probably be distributed over one or several areas which have an extension in longitude as well as along the path.

On the observational side we have drawn attention[9] to the existence of at least one meteor shower whose observed properties are quite sufficient to explain the length of a tektite strewnfield. This is the Cyrillid shower (the great meteor train of 9 February 1913). Although the region over which this shower was observed was very narrow, it is nevertheless helpful in understanding tektite strewnfields, first because the moldavite strewnfield is just about as narrow and secondly because the slight amount of broadening which was observed in this field conforms to the theoretical mechanism of our paper as mentioned above and in fact suggested our mechanism. We find that a satellite which ends its trajectory in an orbit of low eccentricity will be distributed over a long and narrow strewnfield, in contrast to the broadened distribution that is expected in the orbits of high eccentricity.

The comparison which Urey makes between the Cyrillids and the Rochester meteorite is invlid. The Rochester procession was seen over a distance of about 1600 km and was about 6 km in length[10]; it might easily have resulted from the breakup of a normal meteorite.[11] The Cyrillid stream was seen over 10,000 km, and was 1500 km in length; it could only have resulted from the breakup of a natural earth satellite.[12]

On the experimental side, there is now evidence from the distribution of fragments of the MA-6 sustainer (Lieutenant Colonel John Glenn's sustainer) that distribution over large fields is possible. Pieces were found in South Africa over a belt some 850 km long by 100 km wide; and other pieces were found in Brazil along the same orbital path.

In discussing the problem of distribution, Urey raises what we feel to be a most fundamental and interesting question, namely, what happens to the lunar ejecta which does not graze the earth's atmosphere and is not in large blocks? He points out that such material should be far commoner than ordinary tektites. While we do not agree that the probability of a grazing impact is as low as he suggests, for reasons of celestial mechanics which one of us (B. E. S.) has outlined.[13] Urey's main point is undoubtedly correct, namely, that nongrazing encounters should greatly outnumber the grazing encounters. Why, then, do we find tektites but not the other material?

It is risky to try to answer the question, because to do so means a guess at the nature of the tektite parent bodies. Up to this point our argument has been solidly founded on physical principles; but here we must speculate. It seems to us likely, however, from the glassy inclusions

seen in some tektites that the parent bodies are also glassy and possibly slaggy. If so, they might resemble Darwin Glass, a slaggy material found in large quantities over a very limited area in Tasmania. It was accepted as tektite by Suess[14] and others. It may be an impactite, since nickel-iron spherules were reported by Spencer[15] and coesite by Reid and Cohen[16]. It resembles Wabar glass in its physical form.[17] No crater has yet been found in its vicinity.

Another possibility is the Igast object,[18] a slaggy body with a tektite composition which was reported by reliable witnesses to have fallen with the usual meteoritic accompaniments of sound and flash. Igast was generally discredited after Michel[19] attacked it; but O. Schiener has allowed us to examine the hand specimen from which Michel worked. Lowman and O'Keefe[20] found evidence that Michel's hand specimen is unrelated to the witnessed fall; in particular it weighs 10 g more than all the material collected from the fall, so that the possibility of lunar origin remains open.

The fate of the Igast object is perhaps one clue to the fate of the directly falling bodies; nobody believes they are meteorites because their chemistry is wrong.

A second clue is also afforded by Igast; it appears to have contained larger than usual amounts of chlorides. Combined with a porous structure, this would guarantee rapid dissolution. Tektites, on the other hand, have little chloride, perhaps because it escaped in the fusion-stripping stage; and they constitute a solid glass, low in alkali and high in silica, of remarkable durability. Hence the finds are likely to be weighted in favor of tektites. A further point is that solid glass attracts attention as a semiprecious stone; slag does not.

Thus it is possible that among the hundreds of slaggy objects which are annually offered to museum curators around the world as meteorites, a few are genuinely from the sky. We urge that museum curators test these objects with a blowpipe before rejecting them; objects with high melting points, which do not froth when they melt, and do not give off hydrogen sulfide when broken, should be studied for density in the powdered form and index of refraction. If the density is between 2.30 and 2.50 and the index of refraction between 1.47 and 1.52, the objects should be chemically analyzed.

Against the theory of lunar origin the strongest argument is that it makes the source of the tektites remarkably like the earth in its chemical properties. Perhaps, however, the moon really is much like the earth in its chemistry.

REFERENCES AND NOTES

1. Urey, H. C., *Science*, **137,** 746 (1962).
2. ———, *Nature*, **179,** 556 (1957).
3. Shoemaker, E. M., "Interpretation of lunar craters," in *Physics and Astronomy of the Moon*, Z. Kopal, Ed. (Academic Press, New York, 1961), p. 333.
4. Adams, E. W., and Huffaker, R. M., *Nature*, **195,** 681–684 (1962).
5. Taylor, G. I., *Proc. Roy. Soc. London Ser. A*, **201,** 159 (1950).
6. Glasstone, S., Ed., *Effects of Nuclear Weapons* (U.S. Govt. Printing Office, Washington, D.C., 1957).
7. Kopal, Z., *Nature*, **181,** 1457–1458 (1958).
8. Suess, F. E., *Jahrbuch Geol. Reichsanstalt, Wien*, **50,** 193 (1900).
9. O'Keefe, J. A., *Science*, **133,** 562 (1961).
10. Farrington, O. C., *Mem. Natl. Acad. Sci. U.S.*, **13,** 382 (1915). A review of Farrington's report of the Rochester meteorite suggests that the length of 6 km is the result of a numerical blunder. Although the angles which are given are not sufficient to make it possible to redo the calculations it is fairly clear that the length which should have been obtained from them was closer to 60 km than 6 km. It is probable that a decimal error was committed. This weakens the above argument, but does not destroy it.
11. Hoffmeister, C., *Die Meteore* (Akademische Verlagsgesellschaft M.B.H., Leipzig, 1937), p. 78.
12. O'Keefe, J. A., *J. Roy. Astron. Soc. Can.*, **53,** 59 (1959).
13. Shute, B., *Astron. J.*, **67,** 283 (1962).
14. Suess, F. E., *Mitt. Geol. Ges. Wien*, **7,** 51 (1914).
15. Spencer, L. J., *Nature*, **132,** 571 (1933).
16. Reid, A. M., and Cohen, A. J., *J. Geophys. Res.*, **67,** 654 (1962).
17. Spencer, L. J., *Nature*, **131,** 117 (1933).
18. Grewingk, C., and Schmidt, C., *Archiv Naturkunde, Liv-, Ehst- u. Kurlands*, 1st ser. Mineralogical Sci., **3,** 421 (1864).
19. Michel, H., *Ann. Naturhist. Hofmuseums, Wien*, **27,** 1 (1913).
20. O'Keefe, J. A., and Lowman, P. D., *NASA Technical Note D–1151*, abstract in *Geochim. Cosmochim. Acta*, **25,** 158 (1961).

EXPERIMENTAL STUDIES ON BOWEN'S DECARBONATION SERIES, II. *P-T* UNIVARIANT EQUILIBRIA OF THE REACTION:

Forsterite+Calcite=Monticellite+Periclase+CO_2

L. S. WALTER

Goddard Space Flight Center

The *P–T* univariant equilibrium curve for reaction A, calcite+forsterite=monticellite+periclase+CO_2, has been experimentally determined to pass through the points: 725°C, 1300 psi(CO_2); 850°C, 4400 psi(CO_2) and 925°C, 10,000 psi(CO_2). The curve is coincident, within the limits of experimental error, with curves for reactions B and C: 2 calcite+forsterite+diopside=3 monticellite+2CO_2, and calcite+deopside=akermanite+CO_2, respectively. The P_{CO_2}–*T* field of stability of the assemblage monticellite-forsterite-calcite-CO_2 which lies between the curves for reactions A and B is thus very small. The forsterite-calcite assemblage is usually produced by contact metamorphism by granitic rock types. Equilibrium pressure curves for reaction A are derived.

INTRODUCTION

This is the second paper in a series presenting the results of experimental investigation of decarbonation reactions. Considering the effect of rising temperature at constant CO_2 pressure on the phases in the CaO–MgO–SiO_2–CO_2 system, Bowen[1] proposed that, after the formation of monticellite and akermantie by the reactions:

$$2CaCO_3+CaMgSi_2O_6+Mg_2SiO_4=3CaMgSiO_4+2CO_2$$

(calcite+diopside+forsterite=monticellite+CO_2)

and

$$CaCO_3+CaMgSi_2O_6=Ca_2MgSi_2O_7+CO_2$$

(calcite+diopside=akermanite+CO_2)

the next logical step in decarbonation was represented by the reaction:

$$Mg_2SiO_4+CaCO_3=CaMgSiO_4+MgO+CO_2$$

(forsterite+calcite=monticellite+periclase+CO_2).

EXPERIMENTAL PROCEDURE

Many of the experimental techniques were described in the first paper of this series.[2] Reagent-grade $CaCO_3$ and MgO were used as sources of calcite and periclase; monticellite and forsterite were prepared by heating appropriate oxide mixes to 1400°C.

The samples were run in cold-seal pressure vessels using CO_2 as pressure medium and were contained in gold-foil envelopes which were pervious to the gas. In each run both the low and the high temperature assemblages, contained in separate envelopes, were used as starting mixtures.

Reaction rates were too slow to permit complete reaction at all times, except when runs were purposely placed at some distance from the univariant pressure-temperature (P–T) curve for this reaction. Therefore determination of the curve was dependent on recognition of the reaction's beginning. The addition of a small amount of distilled water (about 15–30 mg) to the pressure vessel before the runs increased the rate of reaction considerably. This water represented approximately 3 mole percent of the vapor in the pressure vessel, which has a capacity of 5 cm^3.

Determination of small amounts of monticellite produced by the reaction of forsterite and calcite was made on the basis of the presence of the strong x-ray diffraction line of monticellite at 50.4° 2θ (CuK_α, $d = 1.8105$A). The small amount of periclase produced by reaction could not be detected, probably because it reacted with the vapor to produce magnesite upon quenching. Even when run in the P–T stability in which the monticellite-periclase assemblage was compatible with CO_2, this assemblage yielded a small amount of carbonate upon quenching. When this assemblage reacted, however, a large quantity of calcite was produced. With the use of a petrographic microscope, this was easily distinguishable from quench carbonate by its greater abundance. In most cases, forsterite could be identified as a product of a run by x-ray diffraction techniques.

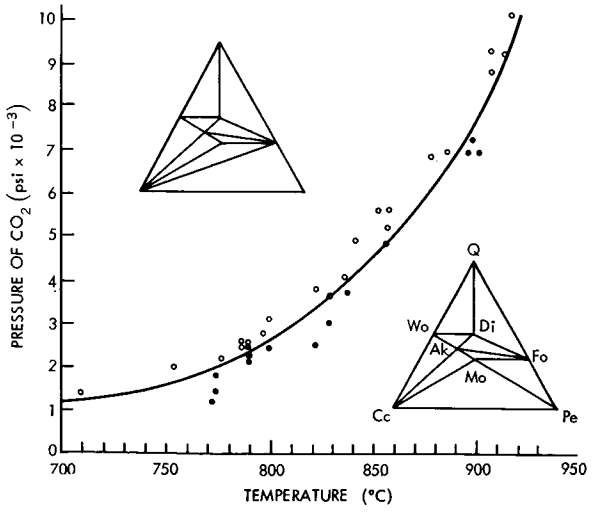


FIGURE 1.—Experimentally determined univariant pressure-temperature curve for the reaction: forsterite + calcite = monticellite + periclase + CO_2.

RESULTS

The univariant P–T curve determined for the "monticellite-periclase" reaction is shown in figure 1, and the data for the critical runs are listed in table 1.

The triangles shown in figure 1 indicate the positions of some of the phases in the compositional tetrahedron CaO–MgO–SiO_2–CO_2. Carbon dioxide is envisioned at the apex of the tetrahedron, which is above the plane of the paper. Those phases which can coexist with the vapor are on, or projected onto, the base and are connected by solid lines. Calcite thus appears at the lower left-hand corner. Three crystalline phases which are mutually stable with CO_2 are shown at the corners of compatibility triangels. This method of projection is more fully described by Bowen.[1]

The experimental error is considered to be ±7°C and ±5 percent of the total pressure. The P–T curve shown in figure 1 is coincident, within the limits of experimental error, with the curves for the formation of monticellite and akermanite reported in the first paper. No explanation can be offered for this relationship.

Because they are located so close together, there is a possibility of the intersection of the P–T curves for the reactions $Fo + Cc = Mo + Pe + CO_2$ and $2Cc + Fo + Di = 3M_0 + 2CO_2$.* The intersection would necessitate an invariant point which is

*The following abbreviations will be employed: Cc = calcite, Fo = forsterite, Pe = periclase, Mo = monticellite, Di = diopside, Q = quartz, and Wo = wollastonite.

TABLE 1.—*Data Used to Determine the Location of the Univariant Pressure-Temperature Curve for the Reaction:* $Cc+Fo=Mo+Pe+CO_2$

Temperature (°C)	Pressure (psi)	Duration (hr)	Starting material	Products identified
710	1400	149	Mo, Pe	Cc
754	2000	144	Mo, Pe	Cc
772	1200	114	Fo, Cc	Mo
774	1450	90	Fo, Cc	Mo
774	1800	117	Fo, Cc	Mo
777	2200	67	Mo, Pe	Cc
787	2600	187	Mo, Pe	Cc
787	2500	162	Mo, Pe	Cc
790	2500	95	Fo, Cc	Mo
790	2500	95	Mo, Pe	Cc
790	2100	119	Fo, Cc	Mo
790	2600	144	Mo, Pe	Cc, Fo
790	2300	87	Fo, Cc	Mo
797	2800	163	Mo, Pe	Cc
800	2450	239	Fo, Cc	Mo
800	3100	114	Mo, Pe	Cc
822	2500	118	Fo, Cc	Mo
823	3800	163	Mo, Pe	Cc
829	3000	64	Fo, Cc	Mo
830	3600	89	Mo, Pe	Cc
837	4000	40	Mo, Pe	Cc
838	3700	117	Fo, Cc	Mo
843	4900	69	Mo, Pe	Cc
854	5600	71	Mo, Pe	Cc
857	4800	42	Fo, Cc	Mo
858	5200	114	Mo, Pe	Cc
859	5600	47	Mo, Pe	Cc
880	6800	65	Mo, Pe	Cc, Fo
888	6900	47	Mo, Pe	Cc
898	6900	114	Fo, Cc	Mo
900	7200	96	Fo, Cc	Mo
905	6100	67	Fo, Cc	Mo
910	8700	45	Mo, Pe	Cc
910	9200	48	Mo, Pe	Cc
917	9100	70	Mo, Pe	Cc
920	10,000	46	Mo, Pe	Cc

illustrated, diagramatically, in figure 2. The spatial relationship among the P–T curves is determined by the compositional relations among the phases involved in the reactions according to principles developed by Schreinemakers.[3] One of the six P–T curves which meet at the invariant point represents the reactions of monticellite and forsterite to form periclase and diopside with decreasing temperature. Several runs at temperatures as low as 465°C and 3000 psi (H_2O) (water did not enter the reaction) showed that monticellite-forsterite was the stable assemblage since it formed by reaction of periclase and diopside. This indicates that the invariant point, if it occurs, will be below 465°C and that at higher temperatures the curve for the reaction $Fo+Cc=Mo+Pe+CO_2$ lies at (slightly) higher temperatures than the curve for the reaction $2Cc+Fo+Di=3Mo+2CO_2$.

DISCUSSION

Critical Assemblages

Assemblages indicative of temperatures below the curve for the reaction $Cc+Fo=Mo+Pe+CO_2$ are monticellite-forsterite-calcite-CO_2 and forsterite-periclase-calcite-CO_2. These assemblages are replaced, at temperatures above the curve, by forsterite-monticellite-periclase-CO_2 and monticellite-periclase-calcite-CO_2.

Forsterite can coexist with calcite, periclase, and CO_2 at temperatures above the curve for the reaction dolomite+calcite=periclase+CO_2,[4] and

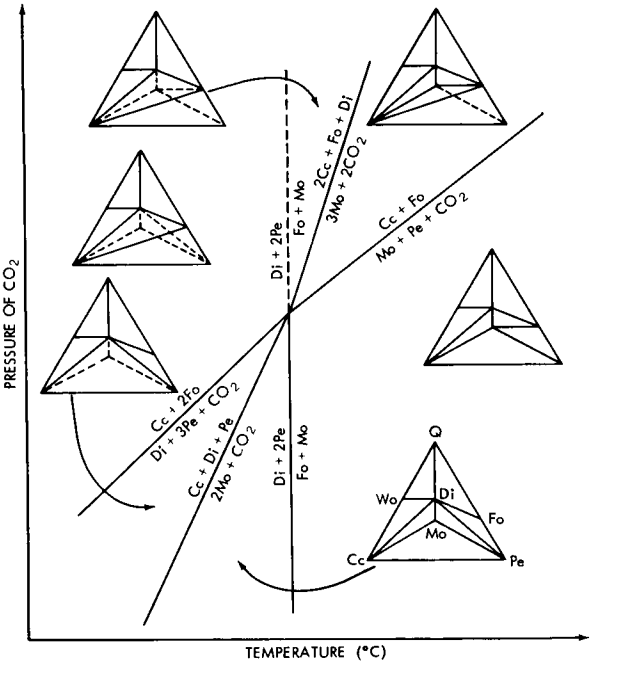


FIGURE 2.—Possible quarternary invariant point involving the phases monticellite, periclase, forsterite, calcite, diopside, and CO_2.

at temperatures below the curve for the reaction $Cc+Fo=Mo+Pe+CO_2$.

The assemblage monticellite-forsterite-calcite-CO_2 is stable only at temperatures above the curve for the reaction $2Cc+Fo+Di=3Mo+2CO_2$. As previously states, the P–T curve for this reaction lies at temperatures slightly below that of the curve for the reaction $Cc+Fo=Mo+Pe+CO_2$. The P–T field of stability of the monticellite-forsterite-calcite-CO_2 assemblage is therefore exceedingly small. Solid solution of fayalite in the olivine, however, may cause the limit of this field to be increased by raising the temperature of the reaction, calcite+olivine=Fe-monticellite+periclase+CO_2.

Equilibrium Pressure Curves

The above results apply only to the condition in which the pressure on the solid phases is equal to that pressure on a pure fluid phase. If, however, three variables, P_f (pressure on the fluid phase), P_s (pressure on the solid phase), and T (temperature) are considered, the univariant P–T curve becomes a divariant surface. This surface may be contoured by curves of constant equilibrium pressure ($P_{E(i)}$). The equilibrium pressure of component i of the fluid is defined by Greenwood[5] as "the pressure of the pure component which is at the same temperature of the system and which has the same chemical potential as the component in the system." This concept was developed by Thompson[6] and Greenwood,[7] and was applied to components of impure fluid phases by Barnes and Ernst,[8] and Walter.[9] It was applied to decarbonation reactions in the first paper.[2]

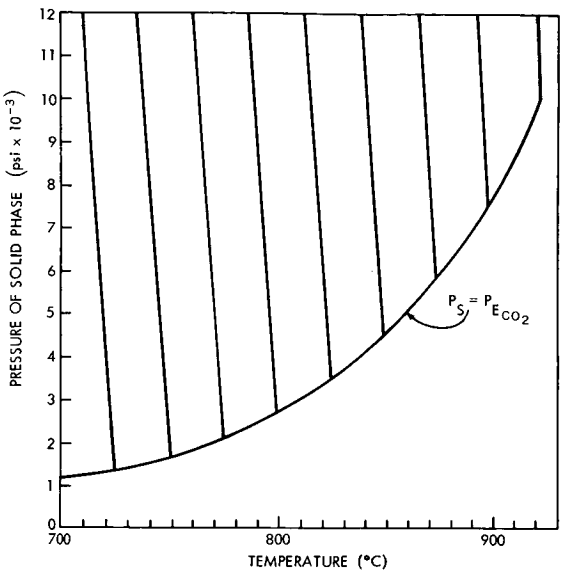


FIGURE 3.—Equilibrium pressure ($P_{E_{CO_2}}$) curves for reaction $Cc+Fo=Mo+Pe+CO_2$.

Curves of constant equilibrium pressure are presented (figure 3) for the reaction $Cc+Fo=Mo+Pe+CO_2$. Steps in the calculation of values for the slope of these curves are given in table 2.

TABLE 2.—*Steps in the Calculation of* $(\partial P_s/\partial T)P_{E_{CO_2}}$.

Temperature (°C)	Pressure (psi)	$\partial P/\partial T$ (psi/°C)	ΔV_{CO_2}* (cm³/mde)	ΔV_{total}† (cm³/reaction)	ΔS_{total} (psi-cm³/°C)	$(\partial P_s/\partial T)P_{E_{CO_2}}$ (psi/°C)
725	1300	10	960	942	9420	−516
750	1600	14	796	778	10892	−597
775	2200	18.5	598	580	10730	−588
800	2600	30	524	506	15180	−832
825	3400	37	414	396	14652	−803
850	4400	49	332	314	15386	−842
875	5800	58	263	245	14210	−779
900	7400	78	217	199	15522	−851
925	10000	147	170	152	22344	−1225

* Reference 10.

† Values for the specific volumes of the condensed phases used in this step were as follows: forsterite, 43.73; calcite, 36.92; monticellite, 51.18; periclase, 11.23 (cm³-gm-mole).

TABLE 3.—*Natural Occurrence of Assemblages Pertinent to the Determined Curve*

Area	Reported assemblage	Igneous rock causing metamorphism	Source
Organ Batholith, New Mexico	Fo–Pe–Cc	Quartz monzonite	Dunham, reference 11
Franklin Mountains, Texas	Cc–Fo–Di	Granite	Lonsdale, reference 12
Skye, Scotland	Fo–Cc	Granite	Tilley, reference 13
Vesuvius, Italy	Mo–Cc–Fo	Extrusive	Tilley, reference 14
Carlingford, Ireland	Cc–Fo–Mo	(Not given)	Osborne, reference 15
Camus Mor, Scotland	Fo–Pe–Cc Mo–Pe	Olivine gabbro	Tilley, reference 16

Natural Occurrence of Pertinent Curves

Naturally occurring assemblages involving the reaction $Cc+Fo=Mo+Pe+CO_2$, as well as the type of igneous rock associated with the metamorphism, are given in table 3. Note that the intrusion of silicic magma generally fails to cause reaction to the high temperature assemblage. On the other hand, metamorphism by an olivine gabbro, in one instance, resulted in the formation of the monticellite-periclase assemblage.

In the reference to the Carlingford district, Osborne[15] states that the igneous rock type which caused metamorphism is obscure. The assemblages he described occur at a locality in which either a basic igneous series or an acidic magma type could have been responsible for metamorphism. In view of its limited P–T stability field, the occurrence of the calcite-monticellite-forsterite assemblage may help to clarify the petrologic relationships in this district.

ACKNOWLEDGMENTS

The guidance of Prof. O. F. Tuttle and the support of the Mineral Industries Experiment Station of the Pennsylvania State University, the National Science Foundation are gratefully acknowledged. The manuscript had the benefit of review by H. L. Barnes, O. F. Tuttle, and P. J. Wyllie.

REFERENCES

1. BOWEN, N. L., "Progressive Metamorphism of Siliceous Limestone and Dolomite," *J. Geology*, **48**(3): 225–274, April–May 1940.
2. WALTER, L. S., "Experimental Studies on Bowen's Decarbonation Series: I—P–T Univariant Equilibria of the 'Monticellite' and 'Akermanite' Reactions," *Am. J. Sci.*, **261**(2): 151–156, February 1963.
3. SCHREINEMAKERS, F. A. H., "Invariant, Monovariant, and Divariant Equilibria," *Proc. Akad. Wetenschappen*, **18**, 116–126, 531–542, 820–828, 1018–1037, 1175–1190, 1384–1398, 1539–1552, 1676–1691, 1915/1916.
4. HARKER, R. I., and TUTTLE, O. F., "Studies in the System CaO–MgO–CO_2, Part 1: The Thermal Dissociation of Calcite, Dolomite and Magnesite," *Am. J. Science*, **253**, 209–224, April 1955.
5. GREENWOOD, H. J., "Water Pressure and Total Pressure in Metamorphic Rocks," *Yearbook* of the *Carnegie Inst.* of Washington, **59**, 58–63, Washington: Carnegie Institution, Washington, Yearbook 1960.
6. THOMPSON, J. B., "The Thermodynamic Basis for the Mineral Facies Concept," *Am. J. Science*, **253**, 65–103, February 1955.
7. GREENWOOD, H. J., "The System $NaAlSi_2O_6$–H_2O–Argon: Total Pressure and Water Pressure in Metamorphism," *J. Geophys. Res.*, **66**(11): 3923–3946, November 1961.
8. BARNES, H. L. and ERNST, W. G., "Ideality and Ionization in Hydrothermal Fluids: The System MgO–H_2O–NaOH," *Am. J. Sci.*, **261**(2): 129–150, February 1963.
9. WALTER, L. S., "Data on the Fugacity of CO_2 in Mixtures of CO_2 and H_2O," *Am. J. Sci.*, **261**(2): 151–156, February 1963.
10. KENNEDY, G. C., "Pressure-Volume-Temperature Relations in CO_2 at Elevated Temperatures and Pressures," *Am. J. Sciences*, **252**(4): 225–241, April 1954.
11. DUNHAM, K. C., "Xenoliths in the Organ Batholith, New Mexico," *Am. Mineralogist*, **21**(5): 312–320, May 1936.
12. LONSDALE, J. T., "Dipyrite and Associated Contact Minerals from the Franklin Mountains of Texas," *Am. Mineralogist*, **14**(1): 26–32, January 1929.

13. TILLEY, C. E., "The Zoned Contact-Skarns of the Broadford Area, Skye: a Study of Boron-fluorine Metasomatism in Dolomites," *Mineralog. Mag.*, **29**(214): 621–666, September 1951.

14. ———, "Contact-Metamorphic Assemblages in the System CaO–MgO–Al_2O_3–SiO_2," *Geol. Mag.*, **62**(734): 363–367, August 1925.

15. OSBORNE, G. D., "The Metamorphosed Limestones and Associated Contaminated Igneous Rocks of the Carlingford District, Co. Louth," *Geol. Mag.*, **69**(815): 209–233, May 1932.

16. TILLEY, C. E., "The Gabbro-Limestone Contact Zone of the Camas Mor, Muck, Inverness-Shire," *Bull. Comm. Geolog. Finlande*, **140**, 97–104, 1947.

VII. PLANETARY ATMOSPHERES

PERCENTAGE CLOUD COVER FROM TIROS PHOTOGRAPHS

ALBERT ARKING

Goddard Space Flight Center

The analysis of approximately 1500 Tiros pictures, on a digital computer, has yielded statistical data on the global distribution of cloud cover from 12 July to 30 September 1961, between 60°S and 60°N. The latitudinal distribution of percentage cloud cover for that period is compared with a fifty-year mean based upon ground observations; the broad features of the two distributions are in agreement.

INTRODUCTION

There is presently a program to study the heat balance of the earth's atmosphere, using the observational data now available on a global scale from the Tiros meteorological satellites.

To understand the dynamical processes in the atmosphere, it is now clear that we must first understand the large scale, global patterns of motion, which serve as the underlying base upon which weather phenomena are superimposed. The large scale patterns are determined by the energy balance—an important aspect being the variations in incoming and outgoing radiation around the globe. The uneven heating of the earth and atmosphere by solar radiation, supplies the primary driving force behind the general circulation.

The net input of radiant energy at any point on the globe is determined by:

1. *The solar flux incident at the top of the atmosphere.* This is determined purely by geometrical considerations, viz., the angle of the sun with respect to the zenith. Aside from seasonal variations, this leads to a cosine law in the latitudinal distribution of solar flux.

2. *The fraction of the solar flux reflected back into space, i.e., the reflective albedo.* This is determined to a very large extent by the cloud cover. The reflectivity of a cloud-covered region can range up to 80%, depending upon the thickness of the cloud. On the other hand, the reflectivity of water and most land terrain varies from about 3 to 15%, the exceptions being desert areas, which may have reflectivities as high as 25%, and snow, which has reflectivities comparable to thick clouds. The average albedo of the earth is 35%, the main contribution coming from the approximately 50% cloud cover which prevails over the earth. The global cloud cover, therefore, causes a significant departure from the simple cosine law in the latitudinal distribution of solar heating and has an important effect on the general atmospheric circulation.

3. *Emission of infrared radiation from the earth and atmosphere.* The infrared emission is determined by the amount and distribution of the infrared absorbing constituents of the atmosphere. In the so-called "window region," 8 to 12 μ, a cloudless atmosphere is essentially transparent, but even a relatively thin layer of clouds is opaque. Hence, the radiation that emanates from the earth

in the 8 to 12 μ region is characteristic of the temperature of the cloud top, or the ground, if there are no clouds. Since the temperature of the cloud top depends upon its height, there is a strong relationship between the energy radiated in the 8 to 12 μ region and the extent and height of the cloud cover.

Although the window contains only a fraction of the total infrared energy, nevertheless, the *variations* of the total infrared energy from time to time and point to point on the globe, are due almost entirely to variations in the 8 to 12 μ region. Therefore, the cloud cover is an extremely important factor in the loss of energy through infrared radiation, in addition to its direct control over the solar energy admitted into the lower atmosphere.

The Tiros meteorological satellites contain equipment for monitoring both the radiation and the cloud cover. Radiometers measure the outgoing radiation from the earth in the infrared and visible regions of the spectrum, while two television cameras photograph the cloud cover.

The analysis and interpretation of the radiation data is presently being pursued at Goddard Space Flight Center, the U. S. Weather Bureau, and at the University of Chicago. A preliminary report of the work at the Goddard Institute for Space Studies was presented last year by Prabhakara and Rasool.[1]

Until now, the cloud cover photographs have been used primarily for the purposes of day-to-day forecasting and for tracking storms and hurricanes. These were the purposes in mind when the camera and television systems were designed.

However, we have been primarily interested in the large scale energy balance of the atmosphere. In view of the important role that cloud cover plays in this problem, we have set out to process the Tiros photographs to obtain statistical data on the global cloud cover distribution. We have analyzed 1447 Tiros III photographs for this purpose, using an IBM 7090 computer. The end product of the analysis is a magnetic tape containing a synopsis of each of the 1447 pictures. By means of a computer program, the tape can be interrogated to give the mean cloud cover over any geographical region during any interval of time.

The Tiros III photographs were taken during daytime from 12 July to 30 September 1961.

FIGURE 1.—Effect of digitization on the quality of a Tiros TV image. (a) A Tiros III photograph taken over the Pacific Ocean southwest of Japan August 21, 1961, is shown as it appears on a video screen. (b) The same picture is displayed using 10 grey levels on an SC–4020 digital cathode ray plotter after digitization into a nominally 500×500 matrix.

The inclination of the satellite orbit restricted the photography to the region between latitudes 60°S and 60°N.

The use of satellite photographs to determine mean cloud cover percentage is based upon the assumption that each picture is an independent random sample of the cloud cover situation. It is

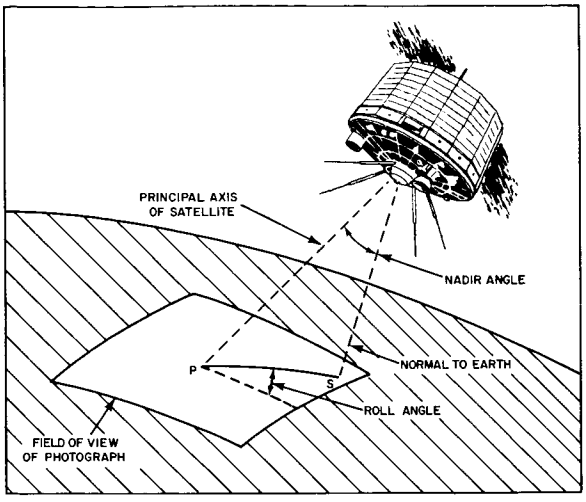


FIGURE 2.—Geometry of Tiros photography. The two television cameras point along the principal or spin axis of the satellite, which is fixed in space. The angle between the principal axis and the normal to the earth, the nadir angle, changes as the satellite orbits the earth. The intersection of the principal axis with the earth is called the principal point, denoted by P, and the point immediately below the satellite, called the subsatellite point, is denoted by S. As the satellite rotates, the image on the vidicon tube appears to rotate about the point P. The angle between PS and a fixed reference line on the vidicon tube defines the roll angle γ, the value of which is required, in addition to the geographic positions of P and S, to transform from picture coordinates to geographic coordinates.

therefore necessary to take cognizance of the distribution of photographs with respect to geographic location and time, when interpreting the results. The fact that pictures within a sequence overlap and are therefore correlated should not invalidate the results, which are based upon treating each picture as an independent sample. The overlapping simply means that regions photographed twice are given twice as much weight in computing averages; this procedure is reasonable because the uncertainty in determining the extent of cloud cover in any region decreases with the number of available photographs of that region.

PROCESSING THE PICTURES

The cloud cover pictures are transmitted from the satellite to the ground station by means of a video system. The image consists of 500 scan lines forming a square pattern on a cathode ray television screen. Photographs of the television image are recorded on film. In addition to the film record, the video signal is recorded on magnetic tape.

To convert the pictures into digital form for insertion into the computer, the taped video signal is sampled at 500 points along each scan line. The results is a 500×500 array of numbers representing the brightness levels in the picture. Brightness is represented by a six-bit code, allowing 64 different brightness levels.

That the quality of the picture is retained after the digitization process, can be seen in figure 1, which compares the original TV image (figure 1a) with the digitized image which is displayed on the SC-4020 digital cathode ray plotter, (figure 1b). Only ten different grey levels were available on the 4020 plotter.

The cameras are oriented with the optical axes parallel to the spin axis of the satellite. Since the satellite is spin-oriented, the direction of the cameras is fixed with respect to space. When the camera is pointed along the normal to the earth, the area viewed is approximately 1000 km square. In general, the camera axis forms an angle with respect to the normal to the earth, called the *nadir angle* (figure 2). Distortions due to the perspective view of the camera and the earth's curvature are properly corrected, taking into account the satellite's position, the spin axis orientation, and the roll angle.

DETERMINATION OF CLOUD BOUNDARIES

The main problem in the analysis of cloud pictures on a computer is the choice of a criterion for distinguishing clouds from clear areas. Re-

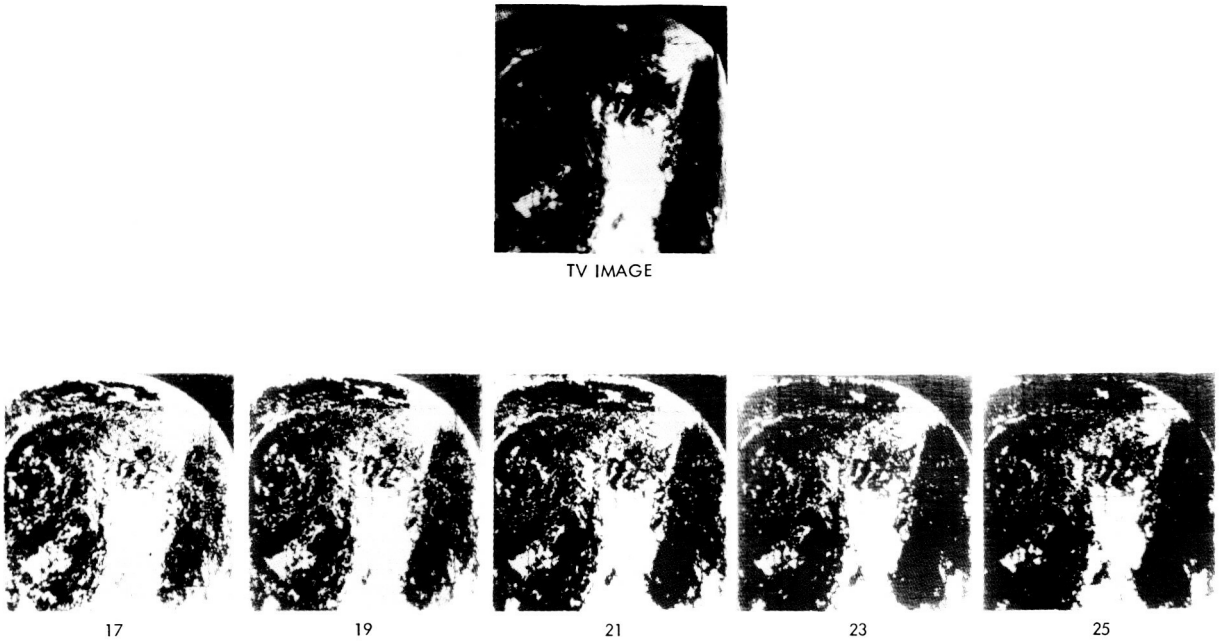


FIGURE 3.—Brightness criteria used in discriminating clouds from clear areas. A series of two-level images, in which intensities greater than a given threshold appear white and all others black, is shown in the bottom row for five different thresholds. The number underneath each two-level image is the brightness level used as the threshold for the discrimination. The original half-tone picture, on top, is compared with each two-level image to find the threshold that most closely reproduces the cloud boundaries that the human observer finds in the half-tone image.

flectivity would be a suitable criterion, for clear areas will have lower reflectivities than clouds except in snow covered regions. The relation between reflectivity and brightness, however, depends upon several factors: the relative angles of the sun and camera with respect to the surface appearing in the photograph, the structure and thickness of clouds, and the nature of the underlying terrain. In addition, there may be variations in the characteristics of the television system during the lifetime of the satellite. The reflectivities of clouds, however, are so much higher than those of the underlying terrain, that the brightness of regions as they appear in a photograph is found to be a reliable criterion. For any one picture it is possible to choose a brightness threshold which distinguishes clouds from other features; furthermore, the threshold is found to vary only slightly from picture to picture within the set of pictures transmitted during any one pass of the satellite over the read-out station. It is therefore necessary to determine a threshold for each read-out orbit. The intent is to choose a threshold which will reproduce the same cloud boundaries that would appear to a human observer studying the picture.

The technique for finding the proper threshold involved use of the SC-4020 cathode ray digital plotter. For a given threshold, a two-level image is produced on the SC-4020, in which grey levels above the threshold are white and other levels are black. For a number of different thresholds, the SC-4020 produces a series of such two-level images—for example, the five images in the bottom row of figure 3. The original picture (figure 1a is reproduced in the upper portion of figure 3) is then compared by the observer with each of the two-level images, and the observer selects the two-level image in which the boundary most closely resembles his impression of the cloud boundaries as seen in the original picture. This establishes the threshold for that picture and the

same threshold is applied to all other pictures transmitted during the same orbit. The list of threshold values for the various orbits forms a threshold table which is used in estimating the cloud cover.

CLOUD COVER PERCENTAGE

Each picture is divided into 121 equal square blocks after eliminating portions of the picture along each border where distortion and noise tend to be excessive. The geographical coordinates of each picture block and the total earth area covered by the block are calculated using the data on the satellite's position, attitude, and roll angle. The distribution of brightness levels—i.e., the fraction of area versus brightness—is determined for each block and entered onto a final magnetic tape along with the geographical data and the time the picture was taken. For each block of every picture that was processed, the final magnetic tape contains latitude, longitude, earth area, time and the brightness level distribution.

Given any geographical region and period of time, the mean percentage cloud cover is determined by scanning the final magnetic tape with a computer program containing the threshold table. All picture blocks that fall into the given geographical region and time interval are used in the analysis. The brightness distribution of each block is summed over all brightness levels above the threshold; the product of this sum with the

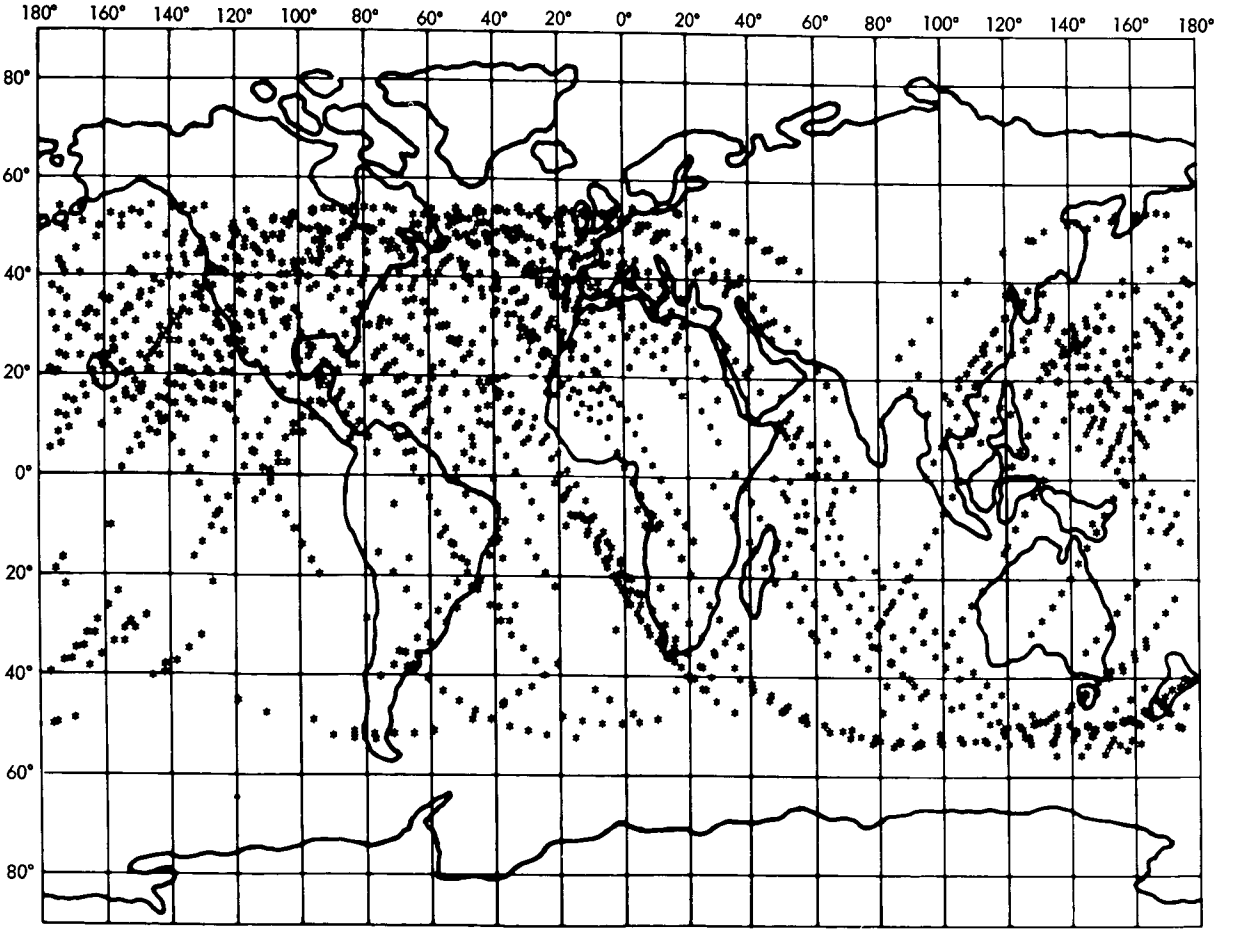


FIGURE 4.—The geographical distribution of the 1447 Tiros III pictures used in the analysis.

total earth area gives the cloud area. The cloud area and the total area of each block are then summed to give the percentage cloud cover.

RESULTS

There were 1447 TIROS III photographs available on video tape which were free from excessive noise and for which we were able to obtain all the data required for the transformation to geographical coordinates. All the pictures were processed on the 7090 computer and are the basis of the results presented here.

The geographical distribution of the pictures is shown in figure 4; each asterisk on the map represents a single photograph. All were taken in the daytime between 12 July and 30 September, 1961 between latitudes 60°S and 60°N. The number of pictures in the Southern Hemisphere is appreciably less than in the Northern Hemisphere and, consequently, errors due to statistical sampling will be greater for southern latitudes. There may also be some statistical bias in northern latitudes due to the sparseness of photographs over Asia.

The average latitudinal distribution of cloud cover is shown in figure 5 for the period 12 July to 30 September 1961. The solid horizontal bars give the mean percentage of earth area covered by clouds in 10° latitude intervals as determined from the TIROS III photographs. The vertical lines passing through the bars indicate the uncertainty estimated to arise in the threshold determination. Possible errors due to statistical sampling are not shown.

The results in figure 5 show that the cloud cover in middle latitudes is the same in the northern and southern hemispheres. However, in tropical latitudes there is an asymmetry, with a local maximum of the cloud cover in the tropics centered at 10°N latitude. This is the average position of the "thermal equator" during the period 12 July to 30 September.

The broad features of the latitudinal distribution of cloud cover obtained from the TIROS photographs are consistent with the known pattern of the general circulation (see, for example, reference 2). Air rising at the thermal equator produces condensation and a relative maximum in the cloud cover, while on the average there is downward motion of cool, dry air at 30°, which explains the relative minimum of cloudiness. The relationship between cloud cover distribution and vertical air currents suggests that the TIROS cloud cover statistics may have an important application in the determination of vertical motions in the atmosphere.

It is of considerable interest to compare the TIROS observations for the summer of 1961 with the climatological distributions of cloudiness found in the literature for the same season. Such distributions have been published by Haurwitz and Austin,[2] Landsberg[3] and others for the globe; and, based on more extensive data, by Telegadas and London[4] for the Northern Hemisphere. The climatological results for the summer season are shown by the dashed histogram in figure 5. The Northern Hemisphere data are taken from Telegadas and London;[4] the Southern Hemisphere data are taken from Landsberg.[3]

The TIROS results for 1961 are seen to be in

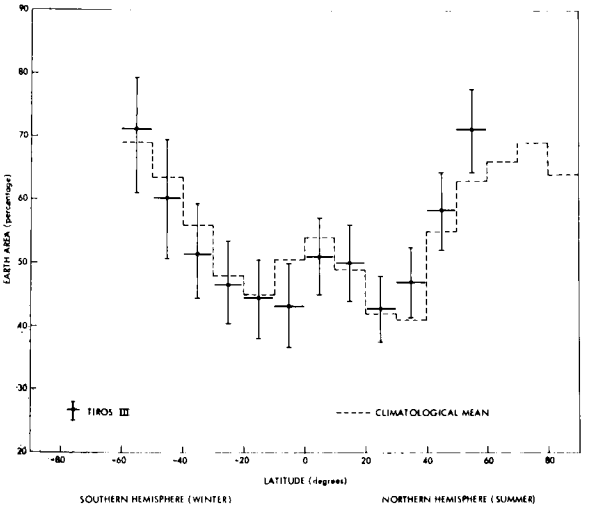


FIGURE 5.—The latitudinal distribution of cloud cover. The solid horizontal bars are the results from Tiros III photographs from July 12 to September 30, 1961; the vertical lines show the estimated errors due to threshold determination. The dashed histogram represents the climatological mean cloud cover based upon ground observations, taken from Telegadas and London (reference 2) for the Northern Hemisphere and from Landsberg (reference 3) for the Southern Hemisphere.

good agreement with the long-term mean of cloud cover distribution obtained from ground observations. The degree of correspondence between our results and the ground based data gives us confidence in this method of analysis of satellite photographs. The availability of more data from subsequent satellites will permit the determination of the geographical distribution of cloud cover over short intervals of time. An increase in the density of observations, coupled with improvements in the techniques for picture analysis presently being developed, should eventually lead to the use of cloud cover pictures automatically analyzed on a computer, to provide the cloud cover distribution as a function of time for use in studies of atmospheric dynamics and long-term changes in climate.

ACKNOWLEDGMENTS

We wish to express our appreciation to Professor Julius London of the University of Colorado for suggesting the original investigation and for his detailed criticism. We are also grateful to Professors J. Charney and R. Goody for helpful suggestions and to Burton Kaufman, John Borgelt, and other members of the staff of Computer Applications, Inc., who developed the programs required for the reduction of the Tiros cloud cover tapes.

REFERENCES

1. PRABHAKARA, C., and RASOOL, S. I., "Evaluation of TIROS Infrared Data," *Proceedings of the First International Symposium on Rocket and Satellite Meteorology*, Washington, D. C., April 23–25, 1962.
2. HAURWITZ, B., and AUSTIN, J. M., "Climatology," New York: McGraw-Hill, 1944.
3. LANDSBERG, H., "Climatology," in *Handbook of Meterology*, F. A. Berry, Jr., et al. New York: McGraw-Hill, 1945, pp. 927–997.
4. TELEGADAS, K., and LONDON, J., "A Physical Model for the Northern Hemisphere for Winter and Summer," College of Engineering, New York University, Research Division, *Scientific Report No. 1*, 1954.

EXPERIMENTAL CONFIRMATION FROM THE TIROS VII METEOROLOGICAL SATELLITE OF THE THEORETICALLY CALCULATED RADIANCE OF THE EARTH WITHIN THE 15-MICRON BAND OF CARBON DIOXIDE

W. R. BANDEEN, B. J. CONRATH AND R. A. HANEL

Goddard Space Flight Center

One of the five channels of the medium resolution scanning radiometer (Bandeen et al., 1961) on the TIROS VII Meteorological Satellite, launched on 19 June 1963, is sensitive within the 15 micron CO_2 band. This measurement was stimulated by the general interest in the infrared horizon of the Earth and was carried out to test earlier theoretical calculations (Hanel et al., 1963). Preliminary data reduction shows good agreement between experimental results and our calculations. The theoretical radiance values of the Earth and their independence of cloudiness, the predicted seasonal effects, and a small amount of limb brightening have been confirmed.

All of the aforementioned effects are of importance in a consideration of the horizon sensing problem. However, a determination of the detailed structure of the horizon, also important to horizon sensing, cannot be made from the experimental data; a much smaller field of view than the 5 deg of the TIROS radiometer would be necessary to resolve its true shape. Of even greater importance, perhaps, is the geophysical implication of the measurements, comprising a global map of the temperatures in the lower stratosphere.

The spectral response of the instrument is shown in Fig. 1. The absorptivity of the thermistor bolometer and the reflectivities of the chopper and prismatically shaped reflector are taken into account, as well as the transmission characteristics of the interference filter and the KRS-5 lens. The effective aperture of the instrument is 0.5 cm^2. The peak transmission of the filter is relatively low and the bandwidth narrower than necessary; both factors contribute to the lower signal-to-noise ratio in this channel compared to the other channels. The average generalized absorption coefficient for carbon dioxide within

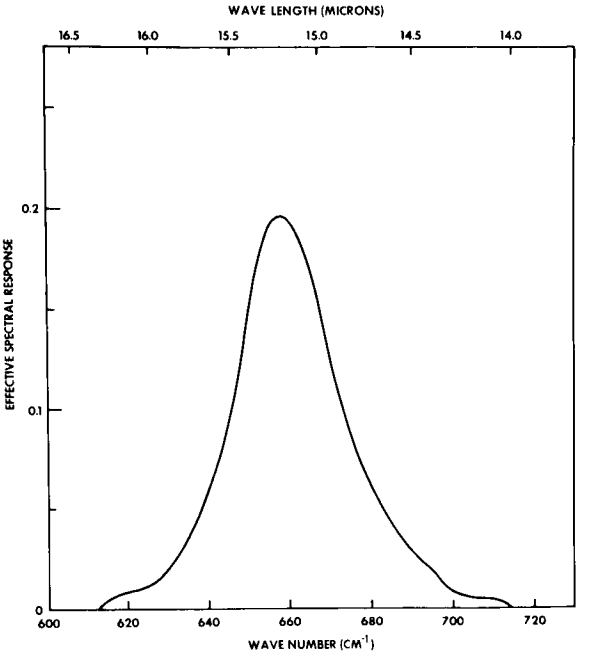


FIGURE 1.—Effective spectral response of the TIROS VII 15-micron channel.

the filter function ($\log L \sim 0.25$) is only slightly higher than the average generalized absorption coefficient used in our calculations. Computed and experimental data are therefore not exactly comparable, but the difference is insignificant for most purposes.

Analog traces of signals from all five channels of the TIROS radiometer are shown in Fig. 2. Six individual scans corresponding to six rotations of the satellite can be identified. Between scans, both sides of the radiometer view outer space, which serves as a zero radiation calibration reference. The calibration of the three channels sensitive to infrared radiation is given in terms of the temperature, T_{BB}(deg K), of a black body filling the field of view of the radiometer. The effective radiant emittance of the target, $\bar{W}$ (watts per meter^2), is calculated by integrating the Planck function at the temperature T_{BB} over the effective spectral response of the channel. The calibration of the two channels sensitive to solar radiation is given in terms of the effective radiant emittance

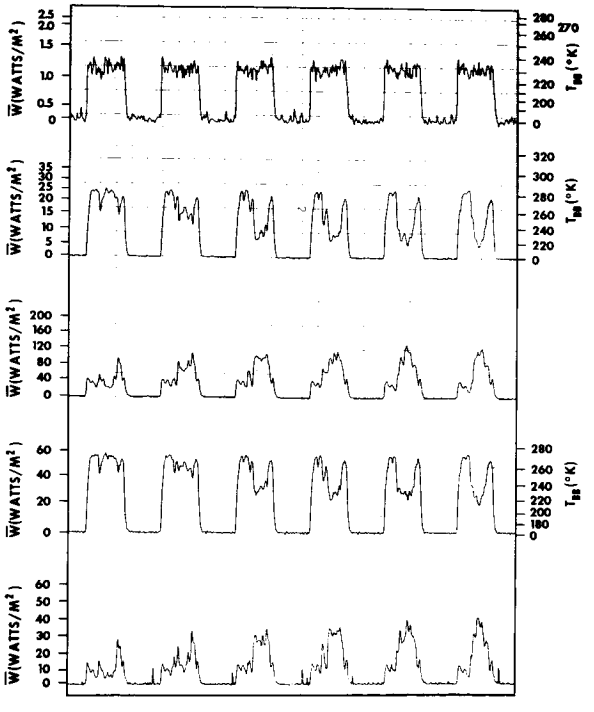


FIGURE 2.—Oscillogram showing six scans off the east coast of Africa by the TIROS VII radiometer. The approximate wave length intervals of the channels from top to bottom are: 14.8–15.5 μ, 8–12 μ, 0.2–5 μ, 8–30 μ, and 0.55–0.75 μ. The effects of high clouds are illustrated.

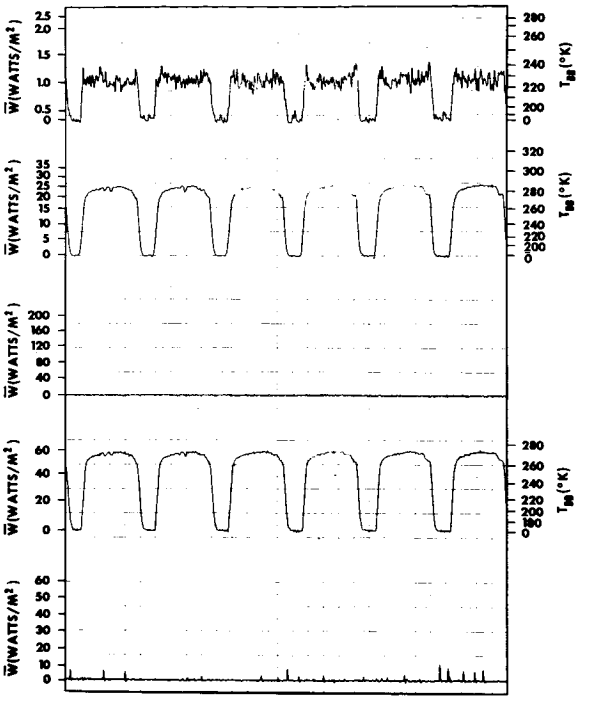


FIGURE 3.—Oscillogram showing six scans over the tropical Pacific Ocean at night. The channels are in the same order as in figure 2. Limb effects are illustrated.

from a diffuse reflector illuminated by the sun. Actually, the 5 deg field of view more nearly measures the effective specific intensity or radiance, $\bar{N}$ (watts per meter^2 per sterad) in the direction of the satellite. Of course, for an ideally diffuse source, the relationship $\bar{W} = \pi \bar{N}$ holds. The low temperatures in the 8- to 12-micron and 8- to 30-micron channels, visible in the middle of each swath, and simultaneous maxima in the channels sensitive to reflected solar radiation, identify clearly an extensive area of high cloudiness. The 15-micron channel shows no correlation with the other channels; the radiance of the earth in this wavelength region is virtually independent of the state of cloudiness. The fine structure on top of the 15 micron trace is mostly detector and amplifier noise and not real, as can be seen from the appearance of the space reference level.

A case of limb-darkening in the 8- to 12-micron and 8- to 30-micron regions, and limb-brightening in the 15-micron band can be observed over an apparently cloud-free area in Fig. 3. Again most of the fine structure must be attributed to noise.

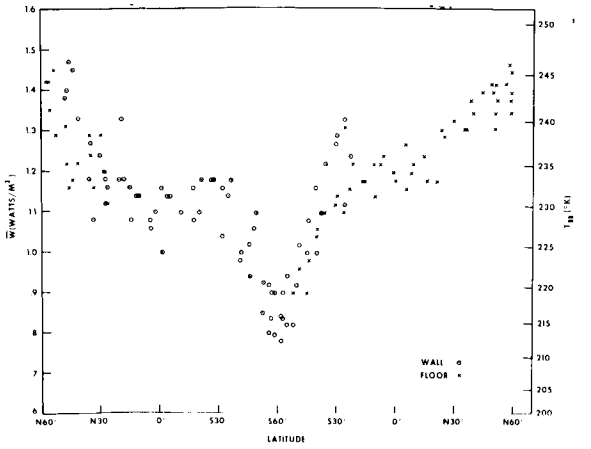


FIGURE 4.—Average radiation levels throughout six orbits vs. the subsatellite latitude. Measurements (made through the satellite baseplate (floor) and side (wall) are separately identified.

In Fig. 4, $\bar{W}$ values averaged over swaths from horizon to horizon, without regard to viewing angle, are plotted versus subsatellite latitude. The values were taken about every fortieth swath (approximately every $4\frac{1}{2}$ min) which is deemed to be sufficient since radiance values change only slowly over the period of one orbit. Data were plotted for the wall and floor sides of the radiometer for six orbits (1, 4, 5, 14, 18 and 19). The lowest values appear over the high latitudes in the southern hemisphere (antarctic winter) and the maxima occur at high latitudes in the northern hemisphere (arctic summer). Data obtained over mid-latitudes and tropical zones exhibit minor variations, but temperatures are generally between 225K and 240K.

The physical meaning of these blackbody temperatures can be judged from Fig. 5, which shows the relative contributions of atmospheric layers to the radiance measured by the radiometer; curves are shown for the 14- to 16-micron region, and the actual filter function of TIROS VII for a zenith angle of zero. Atmospheric emission in the vicinity of 15 to 25 km contributes most, and the regions below 10 and above 40 km contribute only very little, to the radiance at normal incidence. When regions near the horizon are within the field of view of the sensor, the center of mass of the emission function shifts to higher (and warmer) altitudes; hence, limb-brightening occurs. The temperatures shown in Fig. 4 can therefore be interpreted to be average atmospheric temperatures weighted over varying altitudes, depending upon atmospheric structure and sensor viewing angle. The weighting function for the particular case of vertical incidence over a standard atmosphere is indicated by the 14.8–15.5 μ curve in Fig. 5. In theory, using the method proposed by King (1956), the actual temperature profile can be derived from these data. However, the low signal-to-noise ratio will make this a rather difficult task for TIROS VII.

In summary, data from the 15-micron channel

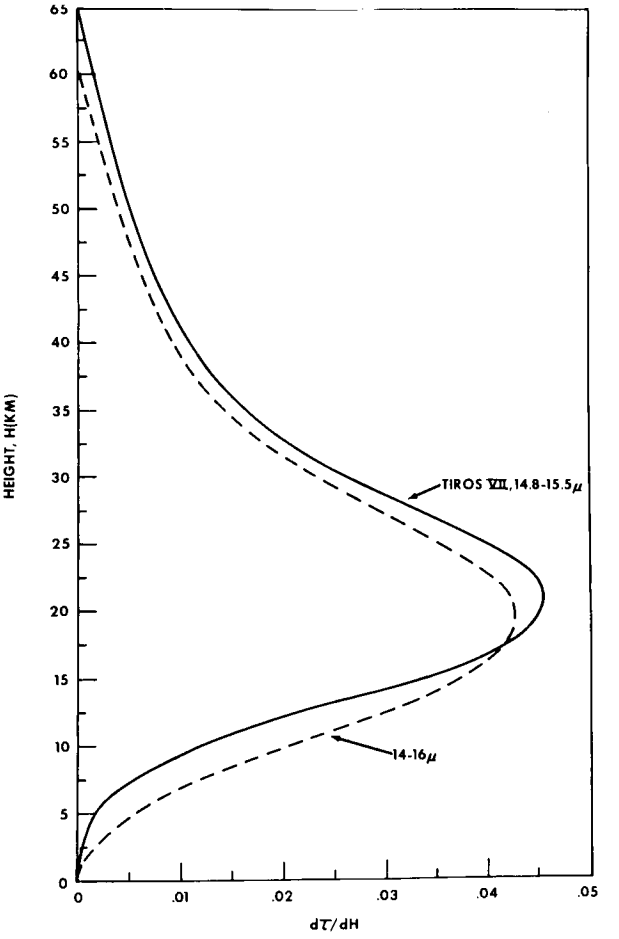


FIGURE 5.—The relative contributions of layers of the ARDC 1959 model atmosphere to the vertical outgoing radiance in the interval 14–16 μ and within the effective spectral response of the TIROS VII 15 μ channel. Since the transmission, τ, from the surface to the top of the atmosphere within these wavelength intervals is essentially zero, the area bounded by each curve is unity.

of the TIROS VII scanning radiometer have confirmed expectations. The minimum and maximum blackbody temperatures observed agree well with theoretically predicted values. Lowest temperatures occur in the Antarctic winter and highest in the Arctic summer. A small but definite amount of limbbrightening is observed. Except for seasonal effects, which presumably can be predicted, the radiance within the 15-micron band is very uniform and virtually independent of cloudiness, a fact most desirable for the purpose of horizon sensing. The observation of the temperature of the lower stratosphere on a global basis may become a valuable tool for the study of the general circulation of the atmosphere. One possible application would be the detection of the "explosive" warming associated with the breakdown of the wintertime polar vortex in the stratosphere.

ACKNOWLEDGMENT

We should like to thank Mr. Andrew W. McCulloch of the Goddard Space Flight Center, who carried out the radiometer experiment on TIROS VII, for kindly making available to us the initial radiation data discussed above.

REFERENCES

BANDEEN, W. R., R. A. HANEL, JOHN LICHT, R. A. STAMPFL and W. G. STROUD, Infrared and reflected solar radiation measurements from the TIROS II meteorological satellite. *J. geophys. Res.*, **66**, 3169–3185, 1961.

HANEL, R. A., W. R. BANDEEN and B. J. CONRATH: The infrared horizon of the planet Earth. *J. atmos. Sci.*, **20**, 73–86, 1963.

KING, J. I. F.: The radiative heat transfer of Planet Earth. *Scientific uses of earth satellites*, Ann Arbor, University of Michigan Press, 133 pp., 1956.

A RADIATION VIEW OF HURRICANE ANNA FROM THE TIROS III METEOROLOGICAL SATELLITE

W. R. BANDEEN, B. J. CONRATH, W. NORDBERG AND H. P. THOMPSON

Goddard Space Flight Center

The Tiros III meteorological satellite (1961 ρ1), containing two television cameras and a family of electromagnetic radiation experiments, was launched on July 12, 1961. Nine days later, the satellite passed directly over Hurricane Anna, the first hurricane of the 1961 Atlantic season. Data gathered by a five-channel medium resolution radiometer during one pass over the Hurricane are presented in the form of maps; and certain implications of the data are discussed. Supporting television pictures are also given.

The design and calibration of the medium resolution radiometer are briefly described.

INTRODUCTION

The Tiros III meteorological satellite (1961 ρ1), launched on July 12, 1961, was the first of the Tiros series whose active lifetime extended over the peak of the Atlantic hurricane season. Two quasi-operational television camera systems carried by Tiros III observed every hurricane of the 1961 season. An experimental medium-resolution radiometer, scanning the earth and its atmosphere in five regions of the electromagnetic spectrum, also gathered data from the same storms from an orbital height of about 780 km.

The design of the radiometers, their calibration, the information flow, and the data reduction techniques have been discussed previously.[1,2,3] Also, the physical signifiance of the experiment and its potential synoptic use have been investigated.[4,5]

In this paper, we shall discuss a sample of automatically processed radiation data gathered by Tiros III over the first tropical cyclone of the 1961 Atlantic season, Hurricane Anna.

DESIGN OF THE EXPERIMENT

To aid in understanding the radiation maps shown, the design and calibration of the radiation experiment should be discussed briefly here. The medium-resolution radiometer flown in Tiros III is practically identical to the first of its kind flown in Tiros II (1960 π1). There are five channels in the radiometer, each made sensitive to a different part of the electromagnetic spectrum by means of filters and other optical elements. The nominal bandwidths of the five channels are given in table 1.

TABLE 1.—*Channel Bandwidths of the Tiros Five-Channel Medium Resolution Radiometer*

Channel	Nominal bandwidth (microns)	Nature of band
1	5.9–6.7	Water vapor absorption
2	7.5–13.5	Atmospheric window
3	0.2–7.0	Reflected solar radiation
4	7.0–32.0	Terrestrial radiation
5	0.5–0.75	Response of the TV system

The radiometer employs a chopper which causes each sensor to view alternately, and at a rapid rate, in two directions 180 degrees apart. The response from each channel is proportional to the

difference in the irradiation of the sensor bolometer from the two directions. The bidirectional axes of the channels are parallel to one another and inclined to the spin (camera) axis by 45 degrees. The satellite spins at about 9.3 rpm, causing the 5 degree instantaneous field of view to scan over the earth and space. The satellite's orbital motion advances the individual scan lines.

When viewing directly downward from a height of 780 km, the radiometer views a "spot" on earth with a diameter of about 68 km. As the nadir angle increases, the scan spot becomes increasingly elongated in the direction viewed. The satellite is spin stabilized, and as it revolves in orbit the radiometer scan geometry over the earth passes through a sequence of rather complicated patterns; however, in any configuration at least one direction of the sensors must always view outer space, which is essentially a zero reference for all five spectral regions. This zero reference is important in calibrating the instrument.

CALIBRATION

The preflight calibration of the three infrared channels is carried out by simulating the space-referenced earth signal in the laboratory. The radiometer's field of view is filled in one direction by a blackbody target at liquid nitrogen temperature (80°K)—which is essentially equivalent to the space reference—while the field of view in the other direction is filled with a blackbody target whose temperature is varied over the range expected when viewing the earth and its atmosphere. Thus, the measurements shown on the infrared maps are in terms of an "equivalent temperature, T_{BB}" of a blackbody filling the field of view which would cause the same response from the radiometer.

In calibrating the solar channels, one direction of the channels is merely masked with black tape to simulate space. To simulate reflected solar radiation from the earth and its atmosphere, a white diffuse reflector of known spectral reflectivity is illuminated at normal incidence from a standard lamp of known spectral intensity at a measured distance away, thus determining the spectral radiant emittance from the reflector. With the illuminated reflector filling the field of view, the output signal from each channel is measured. The spectral radiant emittance from the reflector is integrated over the spectral response curves of the solar channels to yield that portion of the radiation viewed to which each channel responds. The result of such integration is the "effective radiant emittance, $\bar{W}$." Thus, the measurements shown on the reflected solar radiation maps are in terms of that portion of diffuse radiation from a target filling the field of view to which each channel responds, $\bar{W}$. In order to interpret these measurements in terms of reflectance of insolation (or, loosely, "albedo"), we must know the effective radiant emittance, $\bar{W}^*$, which would be measured by each channel if the field of view were filled by a perfectly diffuse surface of unit reflectivity when illuminated by one solar constant at normal incidence. The values of $\bar{W}^*$ for channels 3 and 5 have been calculated to be, respectively, 763.8 w/m^2 and 108.6 w/m^2.

Obviously, any uncertainties in the spectral response curves of the solar channels will cause uncertainties in the subsequent calculations of $\bar{W}$ and $\bar{W}^*$ values. Because of the broad spectral response of channel 3, uncertainties have a much larger effect on this channel than on channel 5. Therefore, we have less confidence in the *absolute* measurements of channel 3 than of channel 5, although the relative measurements for both channels are accurate to within about 3 percent.

CONSTRUCTION OF THE MAPS

The maps shown in figure 1 through 5 were produced by automatic data processing techniques; the shading between contours was done manually to enhance the display. Figure 2b was specially produced at a scale twice as large as in the other maps and the isotherms were hand-traced and labeled to increase resolution in displaying the data.

In producing the maps (except figure 2b), a computer program distributed all data having a sensor nadir angle of 58 degrees or less during the period from 1545 to 1602 GMT, July 21, 1961, over a grid field with a mesh interval between points of 2.5 degrees of longitude (about 278 km square at the equator). Figure 2b was produced in the same way, except that data from 1545 to 1557 GMT were used and the mesh interval between grid points was 1.25 degrees of longitude, thus

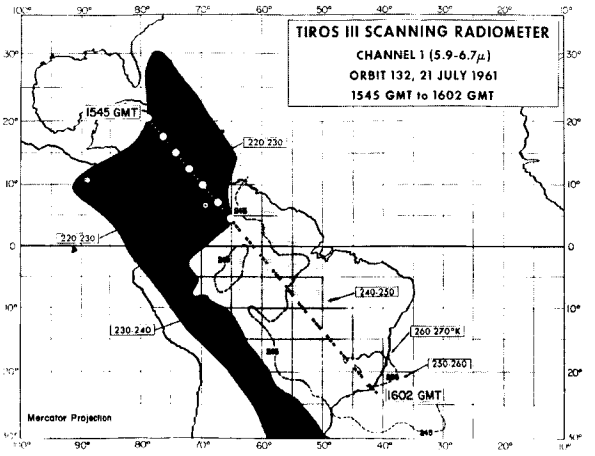


FIGURE 1.—Radiation map from the TIROS III scanning radiometer: channel 1 (5.9 to 6.7 microns), orbit 132 1545–1602 GMT, July 21, 1961. Hurricane Anna is centered at latitude 13.8°N, longitude 72.3°W. Original grid scale: 2.5° of earth longitude per mesh interval. The subsatellite track is shown with subsatellite points indicated by dots for each minute of time.

increasing the resolution by about a factor of four. The population of measurements within a mesh interval varied from 1 along the southwestern edge to more than 60 (or about one-fourth of this number for figure 2b) along the northeastern edge of the data area. This large variance results from the complicated nature of the scan. The minimum spin axis (TV camera axis) nadir angle was 16.0 degrees and occurred at 1558:30 GMT. The subsatellite track is shown on all maps, with subsatellite points indicated by dots for each minute minute of time. The displacement of the data "center of mass" with respect to the subsatellite track is evidence that the camera principal point track lay southwest of the subsatellite track. The computer averages the individual measurements within each mesh interval, accomplishes the contouring of the grid point averages with filler numbers, and prints the maps on a high speed printer. A Mercator latitude-longitude overlay completes the grid print map and locates the data geographically. The maps shown in figure 1 through 5 have been produced by replacing the filler numbers by shading between contour intervals (or tracing the isotherms in figure 2b).

Two limitations of this form data display are re-emphasized: (1) the direction from which a

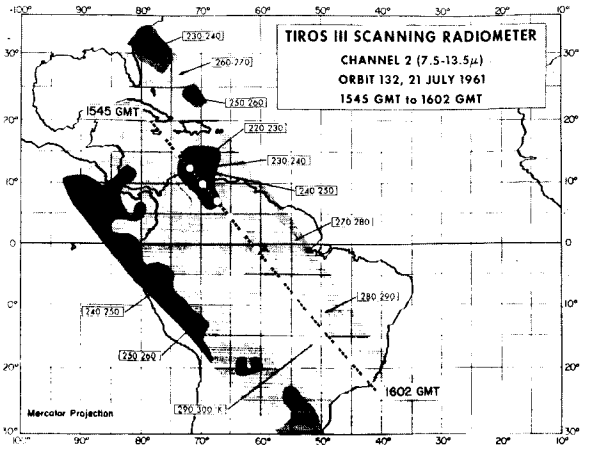


FIGURE 2a.—Radiation map: channel 2 (7.5–13.5 microns). Original grid scale 2.5° per mesh interval.

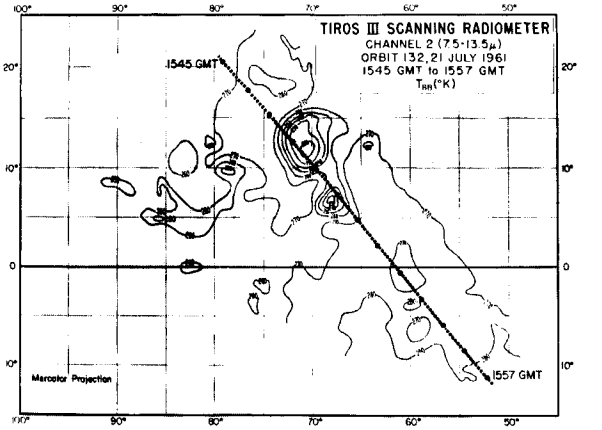


FIGURE 2b.—Detailed map of section of figure 2a. Original grid scale 1.25° per mesh interval.

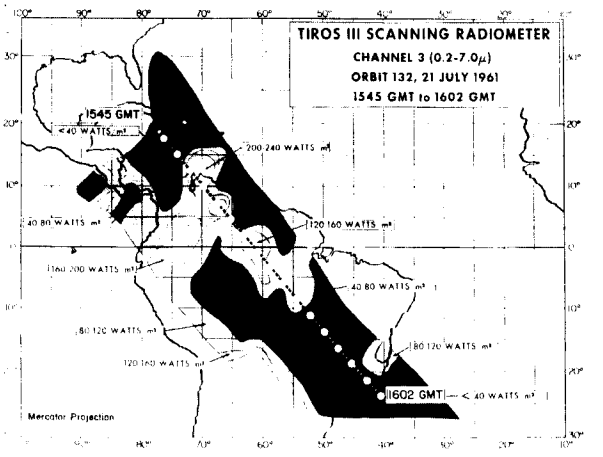


FIGURE 3.—Radiation map: channel 3 (0.2–7.0 microns). Original grid scale 2.5° per mesh interval.

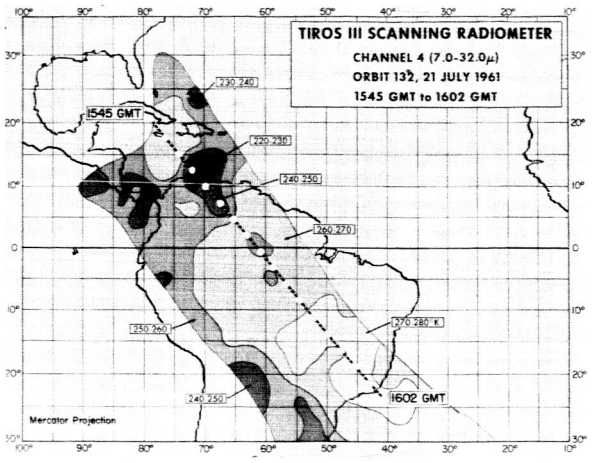


FIGURE 4.—Radiation map: channel 4 (7.0–32.0 microns). Original grid scale 2.5° per mesh interval.

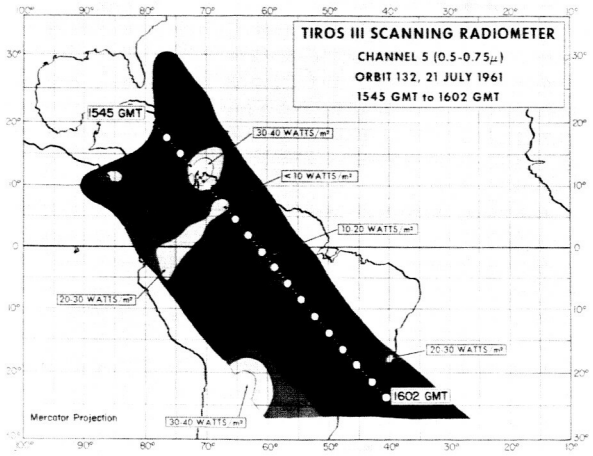


FIGURE 5.—Radiation map: channel 5 (0.5–0.75 microns). Original grid scale 2.5° per mesh interval.

measurement is made (i.e., nadir or zenith and azimuth angles) is not readily apparent, and (2) the number of individual measurements making up each grid point average (or the amount of data smoothing) is not readily apparent. These limitations must be kept in mind when attempting to interpret such maps.

RESULTS

Hurricane Anna moved from a point just north of Venezuela to the coast of British Honduras during the period July 20–24, 1961. At 1550 GMT on July 21, Anna was centered at latitude 13.8°N, longitude 72.3°W, and still growing in intensity. Orbit 132 of Tiros III passed almost directly over the hurricane at that time, and the satellite's attitude was such that all radiometer measurements over Anna were made under sensor nadir angles of less than 15 degrees. Figures 1 through 5 show the stark relief in which the hurricane stands out from its surroundings in all five regions. Three television pictures (figure 6, 7, and 8) taken by Tiros III at about this time show Anna, the

FIGURE 6.—Television picture taken by TIROS III on orbit 132, July 21, 1961, at 1549:50 GMT. Hurricane Anne is the upper cloud mass, centered about 70 miles north of the Guajira Peninsula of Columbia. The Colombian coast runs horizontally to the west and joins Panama (near the upper left fiducial mark). Lake Maracaibo is cradled in the upper right-hand quadrant of the central fiducial cross.

extensive cloudness just to the south of it, and the scattered-to-clear regions over equatorial South America.

Figure 6 shows Anna centered about 70 miles north of the Guajira Peninsula, with the Colombian coastline running laterally to the west and joining Panama. Lake Maracaibo is cradled in the upper right-hand quadrant of the central fiducial cross. A second extensive cloud system is seen south of Anna over Venezuela. The temperatures shown in figures 2 and 4 clearly outline this second system and indicate that its mean top is at a lesser

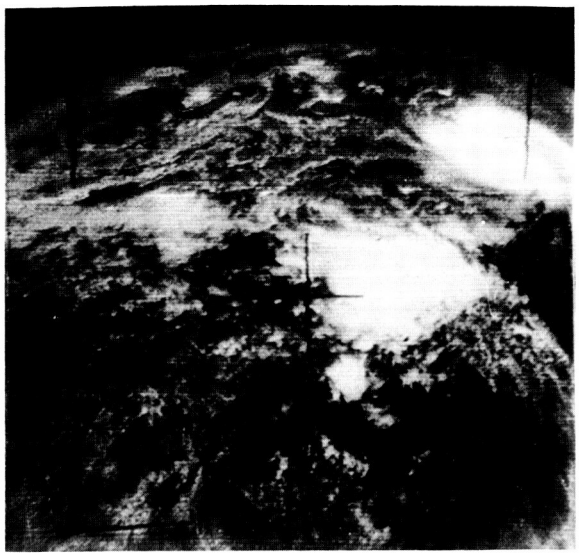


FIGURE 7.—Television picture taken by TIROS III, orbit 132, July 21, 1961, at 1550:50 GMT. The central cloud mass is over Venezuela with Hurricane Anna to the north. The Colombia coast running westward to Panama is seen at the top of the picture.

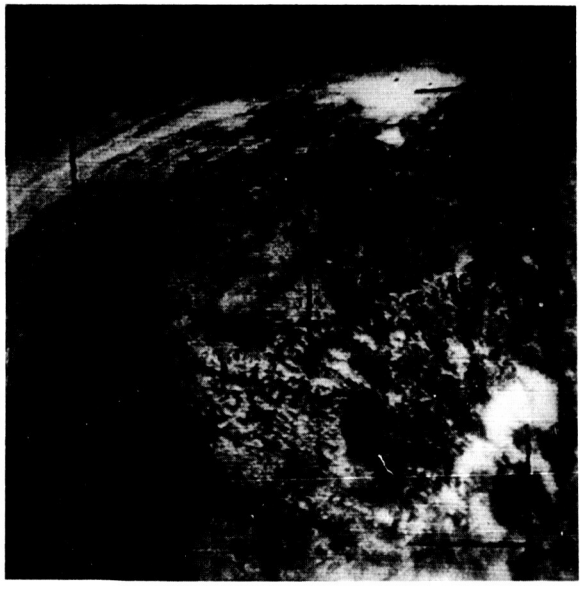


FIGURE 8.—Television picture taken by TIROS III, orbit 132, July 21, 1961, at 1552:20 GMT. The cloud mass which is south of Hurricane Anna is seen at the top of the picture. The central fiducial cross is near the headwaters of the Orinoco River in southern Venezuela. The lower portion of the photograph reveals the scattered-to-clear condition over the Amazon region of Brazil.

height than that of the hurricane. Also, the increased temperatures of these maps over the region stretching from Venezuela to Peru indicate that the band of clouds shown here on the solar maps is at a still lower height, or is thinner, or both.

The hurricane stands out clearly in channels 3 and 5. The actual maximum $\bar{W}$ measurement over Anna from channel 3 was 240 w/m^2 and from channel 5 was 40 w/m^2. The subsolar point at 1550 GMT was at latitude 20.4°N, longitude 55.9° W, some 18 degrees from the storm center. Taking the solar radiation incident upon the hurricane as $\bar{W}^*$ cos 18°, the maximum reflectance r of sunlight in the direction of the satellite is

for channel 3:

$$r = 240/(763.8 \cos 18°) = 33.0 \text{ percent};$$

for channel 5:

$$r = 40/(108.6 \cos 18°) = 38.7 \text{ percent};$$

The maximum r measured by channel 3 is 5.7 percent lower than at measured by channel 5. As was mentioned earlier, measurements by channel 3 might be somewhat in error because of the lack of an accurate measurement of its spectral response. This fact might very well be responsible for the observed difference in r. Of course, this discrepancy does not affect the validity of the relative radiation levels and, therefore, the ability of the data to map cloud systems.

The patterns in the maps of the solar channels are very similar. The large cloud system south of the hurricane over Venezuela and the band of cloudiness stretching from it southwest across Columbia into Northern Peru stand out clearly in figures 3 and 5, and can be identified in the television pictures in figures 7 and 8. The high reflectance region around latitude 22°S and longitude 63°W implies the presence of clouds, but the relatively high temperatures measured by channel 2 indicate that the cloud tops were not very high. The broad band of high channel 2 temperatures running from the Amazon basin across Brazil to the Atlantic Ocean compares well with the low solar channel measurements, implying that there are few clouds and that the effective height of radiation of the window channel is near the surface. The channel 4 and channel 1 maps also

show a similar pattern, and the photograph in figure 8 reveals the scattered-to-clear condition over the Amazon region of Brazil.

The region north of latitude 30°N off Florida is mapped only by channel 2, where values of T_{BB} below 240°K are shown. Widespread cumulonimbus clouds with anvil tops were reported in this region at the time, and undoubtedly contributed to the low temperatures observed. However, a complication in interpreting measurements in this region is that they were all made at nadir angles of nearly 58 degrees. The comparable zenith angles at the top of the atmosphere were about 72 degrees, meaning that the radiometer was viewing through more than three atmospheres. Under these conditions, limb darkening effects become appreciable, even in the window channel, and caution must be used in interpreting the data.

Actually, each infrared channel "sees" many different layers in the atmosphere. The contribution of each atmosphreic layer to the integrated response of a particular channel depends upon the vertical distributions of temperature, pressure, and the various infrared absorbers (primarily H_2O, CO_2 and O_3); and upon the positions and strengths of the various absorption bands with respect to the spectral response curve of the particular channel. Because of strong absorption by water vapor in the 6.3 micron band, generally (in the absence of clouds) a broad region throughout the middle and upper troposphere predominates in contributing to the response of channel 1, with little or no contribution from the surface of the earth.

Because of only weak absorption by ozone and water vapor in the atmospheric window, radiation from the surface of the earth or from clouds predominates in the response of channel 2, with small contributions from the atmosphere. Because channel 4 includes the window as well as strong water vapor and carbon dioxide absorption bands, it receives important contributions from both earth and atmosphere. The air temperature generally decreases with height in the troposphere where a preponderance of the important absorbers is found; hence, it follows that usually the equivalent blackbody temperatures T_{BB} measured by channel 1 should be lowest, those measured by channel 2 highest, and those measured by channel 4 somewhere between the first two. This pattern generally prevails throughout figures 1, 2, and 4, with T_{BB} highest in channel 2, 20 degrees lower in channel 4, and 20 to 40 degrees lower in channel 1. The one exception to this pattern is over the hurricane, where the T_{BB} values for all channels are indicated as 200 to 230°K.

If we define a fictitious level, called the "effective height of radiation," as the height at which the T_{BB} measured by the satellite corresponds to the real atmospheric temperature, it follows from the previous paragraph that the effective height of radiation for channel 1 should be highest and that for channel 2 lowest, with the channel 4 height falling in between. Fritz and Winston[5] have shown that the effective height of radiation from the window channel of Tiros II affords a good estimate of the heights of the tops of large-scale overcast clouds, although cirrus are not clearly defined because they are partially transparent.

From radiosonde data taken near the hurricane, the ambient air temperature ranged from 230 to 220°K in the height interval from about 11,000 to 12,100 meters implying that the smoothed mean height of the hurricane clouds as mapped in figure 2 lay within this layer. However, the smoothing contained in the grid point averages and the 10°K contour intervals used in the maps obscure much of the fine structure which is possible within the 5 degree field of view.

For example, a computer listing of the individual data points making up the maps revealed three individual spots (a spot diameter on earth is about 68 km) over the hurricane where there was a dramatic reversal of the usual order of T_{BB} measurements from the infrared channels coincident with the minimum observed channel 2 temperatures. These three spots, each from a different swath, yielded the values of T_{BB} shown in table 2. The temperature and height of the tropopause averaged from 1200 GMT radiosonde runs made at Kingston, Jamaica, Trinidad, and Swan Island were about 199.5°K and 16,000 meters, respectively. The estimated absolute accuracy of channel 2 is ±5°K. Therefore, the minimum T_{BB} for channel 2 was probably between 199.5° and

TABLE 2.—*Reversal of the Usual Order of Infrared Channel Temperatures in Three Spots Over Hurricane Anna*

Channel	Nominal bandwidth (microns)	T_{BB} (°K)		
		Spot 1	Spot 2	Spot 3
1	5.9–6.7	225	228	228
4	7.0–32.0	219	212	215
2	7.5–13.5	205	207	201

206_TK (i.e., 201°+5°K), and its effective radiating height was probably within 1500 meters of the tropopause. It can therefore be inferred that the highest overcast clouds were within about 2000 meters of the tropopause. The reversal of the usual pattern of T_{BB} measurements can be explained qualitatively by considering an upward translation of the usual effective radiation height sequence (i.e., that for channel 1 highest and for channel 2 lowest) such that the effective radiation height for the window channel is at or near the tropopause, with the others located above the inversion where temperature increases with height.

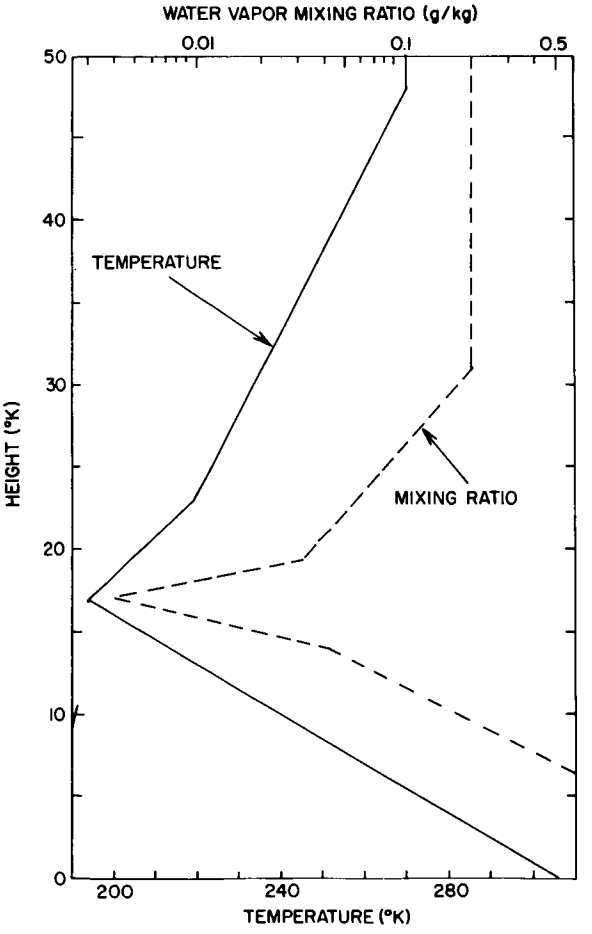


FIGURE 9.—Temperature and water vapor mixing ratio profiles assumed in the theoretical calculations attempting to explain the "reversed" relative behavior of channels 1 and 2 over tropical storm systems.

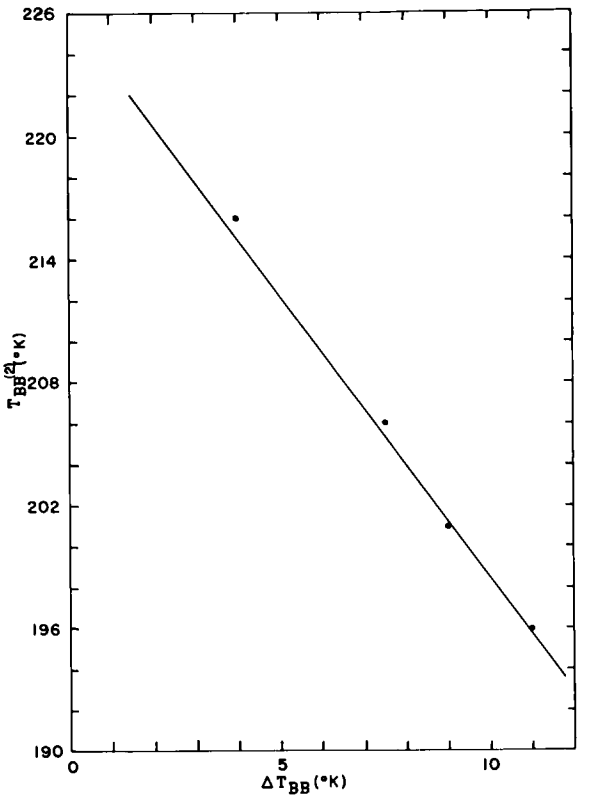


FIGURE 10.—Results of the theoretical calculations. Cloud top temperature, which is assumed to be that seen by channel 2, is shown as the ordinate (channel 2 T_{BB}). The abscissa, ΔT_{BB}, is the difference between the effective blackbody temperature seen in the window region and that seen in the water vapor absorption region.

The estimated absolute accuracy of channel 1 is ±5°K. Therefore, the observed channel 1 T_{BB} corresponding to the minimum observed channel 2 T_{BB} might have been as low as 228°−5°=223°K. It seems difficult to justify sufficient water vapor in the stratosphere to account for a difference in channel 1 and channel 2 T_{BB} measurements of

even 17°K (i.e., 223°–206°). However, the following preliminary calculations, assuming plausible temperature and water vapor profiles for the tropical stratosphere (figure 9), do indicate that differences of at least 10.5°K may indeed be possible, indicating the state of the atmosphere above tropical storms. Reasonable variations about the temperature and water vapor profiles chosen and in the theoretical assumptions regarding absorption could probably yield differences of as much as 17°K.

The specific intensity emerging from the "top" of the atmosphere I is given by the solution of the radiative transfer equation, which may be written in the form

$$I = I_0\tau_0 + \int_{\tau_0}^{1} B[T(u)]\, d\tau(u)$$

where I_0 is the specific intensity at the bounding surface (ground or cloud as the case may be), B is the Planck function, and τ is atmospheric transmission above a level at which the optical depth is u. If τ and the temperature are known as functions of u, then this calculation can be carried out.

In practice, the integral can be evaluated by means of radiation diagrams. A diagram of the Möller type was employed in this particular work

FIGURE 11.—Photograph taken over northwest Africa by TIROS III at about 1309 GMT on July 17, 1961. The African coastline and the Atlantic Ocean can be seen in the upper left corner.

TABLE 3.—*Reversed Data Points in Two Spots Over a Cloud System Shown in Figure 11*

Channel	Nominal bandwidth (microns)	T_{BB} (°K)	
		Spot 1	Spot 2
1	5.9–6.7	213	211
2	7.5–13.5	201	205

This diagram was for a spectral interval of from approximately 6.33 to 6.85 microns, while the channel 1 spectral range of the radiometer was about 6.0 to 6.6 microns. However, the calculations based on the 6.33 to 6.85 micron range should serve to indicate whether the differences observed between channels 1 and 2 are of the right order or not.

Calculations were carried out for cloud tops at and slightly below the trapopause height. The results are shown in figure 10 in which the cloud top temperature as seen by the window channel is plotted against the difference between the temperature seen in the window and that seen in the water vapor absorption region.

Only one other case of a channel 1-channel 2 temperature reversal has been noted in all of the Tiros III radiation data analyzed to date, although it has not been possible yet to survey comprehensively all data and rule out still other occurrences of this phenomenon. This case was associated with a storm system shown in figure 11 which is a picture taken over northwest Africa on July 17, 1961. The African coast line and the Atlantic Ocean can be seen in the upper left corner.

When the cloud system just above the central fiducial cross in the photograph was scanned by the radiometer, the "reversed" data points shown in table 3 taken from two different swaths were found. While the temperature differences seen in these data points are smaller than those associated with the hurricane, they are still probably real and serve to further substantiate the phenomenon of reversed relative behavior of channels 1 and 2. Obviously, much more work is required in this area, and we hope to be able to report additional information at a later time.

CONCLUSIONS

The radiation maps of Hurricane Anna show that a severe storm of this type stands out clearly in all three infrared channels, and, in the daytime, also in the two solar channels of the Tiros medium resolution radiometer. The synoptic use potential of such radiation data in future meteorological satellites seems clear for complementing and augmenting television cameras in the identification and tracking of severe weather systems.

The extraordinary reversal of the usual order of equivalent blackbody temperatures over the hurricane, together with the low channel 2 measurement, suggest that the cloud tops over Hurricane Anna were very high—probably near the height of the tropopause. The magnitude of the difference between the equivalent blackbody temperatures measured by channel 1 and channel 2 infer a very wet stratosphere with a sharp temperature inversion above the cloud tops. However, much more work in this area remains to be done before definite conclusions can be drawn.

ACKNOWLEDGMENTS

The authors wish to express their thanks to Mr. Robert Hite, of Goodard Space Flight Center, who wrote the primary radiation data reduction program, and to Mr. William Callicot, United States Weather Bureau, who wrote the mercator mapping program used in the automatic computer processing of the radiation data given in this paper.

REFERENCES

1. Bandeen, W. R., Hanel, R. A., Licht, J., Stampfl, R. A., and Stroud, W. G., "Infrared and Reflected Solar Radiation Measurements from the Tiros II Meteorological Satellite," *J. Geophys. Res.* **66**(10): 3169–3185, October 1961.
2. Nordberg, W., Bandeen, W. R., Conrath, B. J., Kunde, V., and Persano, I., "Preliminary Results of Radiation Measurements from the Tiros III Meteorological Satellite," *J. Atmosph. Sciences* **19**(1): 20–30, January 1962.
3. Hanel, R. A., and Stroud, W. G., "The Tiros II Radiation Experiment," *Tellus* **13**(4): 486–488, November 1961.
4. Hanel, R. A., and Wark, D. Q., "Tiros II Radiation Experiment and its Physical Significance," *J. Opt. Soc. Amer.* **51**(12): 1394–1399, December 1961.
5. Fritz, S., and Winston, J. S., "Synoptic Use of Radiation Measurements from Satellite Tiros II," *Monthly Weather Review* **90**(1): 1–9, January 1962.
6. Gutnick, M., "How Dry is the Sky?" *J. Geophys. Res.* **66**(9): 2867–2871, September 1961.
7. Nordberg, W., and Stroud, W. G., "Seasonal, Latitudinal, and Diurnal Variations in the Upper Atmosphere," NASA Technical Note D-703, April 1961.

NOTE ON THE THICKNESS OF THE HELIUM ION LAYER*

S. J. BAUER

Goddard Space Flight Center

On the basis of recent experimental results as well as theoretical considerations of the temperature dependence of the light constituents (hydrogen and helium) in the upper atmosphere, a model of the helium ion belt—the "heliosphere"— is constructed. The thickness of the helium ion layer varies significantly with atmospheric temperature: about 2000 km at 1600°K and only about 200 km at 600°K. Correspondingly, charged particle profiles in the topside ionosphere may show a slope corresponding to He^+ at high temperature, but not at low temperatures when the thickness of the helium ion layer is comparable to or less than the scale height of helium ions.

Experimental evidence, in accordance with Nicolet's suggestion,[1] for the presence of helium ions in the upper atmosphere, is now available from ion and electron density profiles[2,3] ion probe retarding potential experiments[4,5] and direct measurements with an RF ion spectrometer.[6] Hanson,[2] who first inferred the presence of He^+ from an experimental ion-density profile obtained by Hale,[7] has called the layer of ionized helium the "heliosphere". In his analysis of Hale's data, Hanson concluded that the thickness of the helium ion layer, extending from 1200 km to about 3400 km, is of the order of 2000 km and that the measured scale height for helium ions corresponded to an atmospheric temperature of 1600°K.

Because of the presence of He^+, the upper ionosphere has to be considered a ternary ion mixture (O^+, He^+ and H^+). Although the diffusive equilibrium distribution of a minor ion species in an ion mixture is influenced by the other constituents—unlike neutral ones, because of the electric field resulting from the slight charge separation between electrons and positive ions,[8–11] the relative concentrations of ionic species behave like neutral constituents.[12] The boundaries of a region where an ionic species predominates can be defined by the transition or equal-concentration levels, i.e. the levels where the concentration of this species is equal to that of the neighboring ionic species.

The equal-concentration level h_{ij} can be expressed in terms of the geopotential height parameter:[12]

$$z_{ij}' = H_{ij} \ln \eta_{ij}, \tag{1}$$

where $z_{ij}' = h_{ij}' - h_0'$ is the geopotential distance between the equal-concentration level and the reference level h_0' at which $[X_i^+]/[X_j^+] = \eta_{ij}$, and $H_{ij} = kT/(m_i - m_j)g_0$ with k the Boltzmann constant, T the absolute temperature, m_i and m_j the masses of the ionic species X_i and X_j and g_0 the acceleration of gravity at the earth's surface. The geopotential altitude h' is related to the geometric altitude h by

$$h' = \int_0^h \frac{g(h)}{g_0}\, dh.$$

In this discussion the subscripts 1, 2 and 3 will refer to O^+, He^+ and H^+, respectively.

The actual situation is complicated by the fact that the relative ion concentrations at the reference level η_{ij} are also functions of temperature. This is particularly true for the ratio η_{31} since the neutral hydrogen concentration at a constant

* The material contained in this report also appeared in *Nature*, January 5, 1963, under the title, "Helium Ion Belt in the Upper Atmosphere."

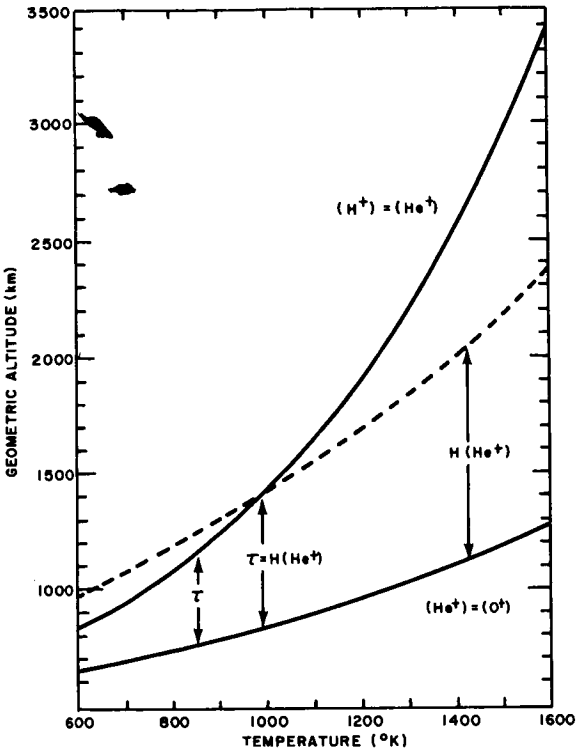


FIGURE 1.—Equal concentration levels $[He^+]=[O^+]$ and $[H^+]=[He^+]$ on a geometric altitude scale versus atmospheric temperature. τ represents the thickness of the helium ion layer; the local scale height for He^+ is indicated by the distance between the dashed curve and the $[He^+]=[O^+]$ curve. The equal-concentration levels are based on the temperature-dependent ion concentration ratios η_{ij} at 500 km given in the text.

altitude depends upon its escape rate from the upper atmosphere. Since this rate increases with temperature, the concentration of hydrogen at a constant level *increases* with *decreasing* temperature. The ratio [H]/[O] at an altitude of 500 km (which we shall adopt as reference level h_0) increases by a factor of 100 for a temperature decrease from 2000 to 1000°K.[13,14] If the proton concentration is governed by the charge exchange equilibrium condition,[13,15] the ratio η_{31} varies with temperature in the same proportion as the ratio of the neutral constituents. The same argument applies if the protons are a result of photoionization rather than charge exchange. Similarly, a proportional variation of [He]/[O] and η_{21} with temperature can be assumed. The helium concentration at 500 km varies only slightly with temperature[14] so that the ratio η_{21} will not vary as strongly as η_{31} with temperature.

For the present model the following values have been adopted: The concentration ratio $[H^+]/[O^+]$ at 500 km varies from $\eta_{31}=10^{-4}$ at 1600°K to $\eta_{31}=10^{-2}$ at 600°K: and the ratio $[He^+]/[O^+]$ at 500 km varies from $\eta_{21}=5\times10^{-3}$ at 1600°K to $\eta_{21}=5\times10^{-2}$ at 600°K. The temperature range corresponds to the extremes at the present level of solar activity. The chosen variation with temperature of η_{31} is in good agreement with Bates and Patterson,[13] and that for η_{21} is consistent with the variation of the helium concentration of Nicolet and Kockarts.[14]

Figure 1 shows that h_{12} and h_{23} transition levels, where $[He^+]=[O^+]$ and $[H^+]=[He^+]$ respectively, as a function of atmospheric temperature

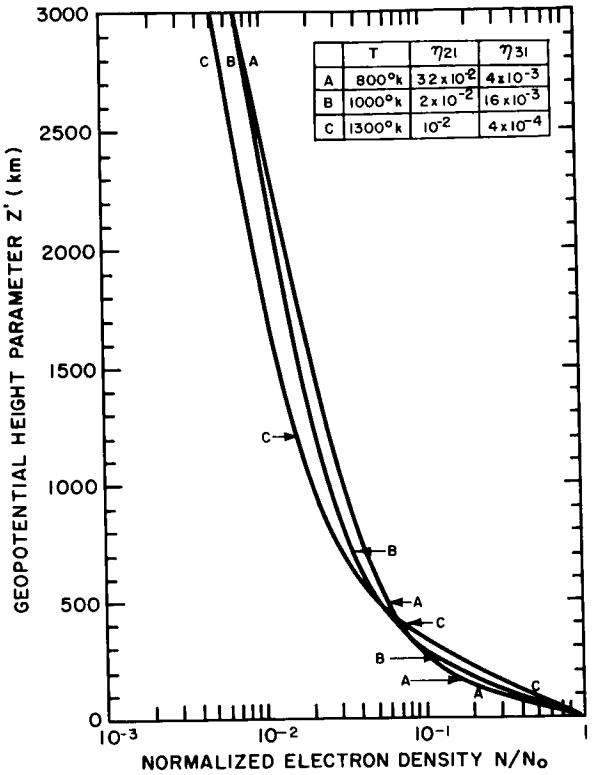


FIGURE 2.—Normalized charged particle density profiles for a ternary ion-mixture corresponding to the listed ion concentration ratios η_{ij} and atmospheric temperatures as a function of geopotential distance z' above the reference level. The solid arrows indicate the equal concentration level $[He^+]=[O^+]$ and the open arrows $[H^+]=[He^+]$.

for the assumed variation of η_{ij} at 500 km, the reference level. The altitude difference $h_{23}-h_{12}$ between the two equal-concentration levels may be called the thickness τ of the helium ion layer. It is obvious that this thickness is drastically reduced as the temperature decreases. This possibility has previously been pointed out for the case where η_{ij} is constant (references 12 and 16).*

If the thickness τ is less than the local scale height for helium ions, $\tau < H(He^+)$; i.e., if the two equal-concentration levels h_{12} and h_{23} occur within one scale height, no distinct slope corresponding to He^+ will be distinguishable in an ion or electron density profile. This is the case for temperatures of the order of 1000°K or less. Figure 2 which presents normalized charged particle density profiles based on a ternary ion mixture for 800, 1000 and 1300°K and the corresponding $\acute{e}_{ij}$ versus the geopotential distance z' above the reference level, illustrates the fact that identification of helium ions from profiles is virtually impossible for low temperatures. The profiles shown in figure 2 were computed from the following formula:[12]

$$N(z') = N_0 \exp\left(-\frac{1}{2}\left\{\left(\frac{z'}{H_1}\right) - \ln\left[1 + \eta_{21}\exp\left(\frac{z'}{H_{12}}\right) + \eta_{31}\exp\left(\frac{z'}{H_{13}}\right)\right] + \ln\left(1+\eta_{21}+\eta_{31}\right)\right\}\right), \quad (2)$$

where N and N_0 are the electron (or total ion) density at z' and at the reference level $z'=0$ respectively and all other symbols have their previously defined meaning.

A recent nighttime electron density profile reported by Ulwick and Pfister,[17] failed to show evidence for the presence of He^+. In the light of the foregoing discussion, failure to detect He^+ from a profile does not prove the absence of He^+ but can be explained by the fact that the thickness of the helium ion layer at that time was comparable in magnitude to the scale height of He^+.

In fact, with an ion trap, Donley[5] recently obtained a nighttime ion density profile from which the presence of helium ions is not obvious. However, a preliminary retarding potential analysis of the same data indicates the presence of all three ionic constituents (O^+, He^+ and H^+). The combined ion density and composition data are consistent with the concept of a thin helium ion layer at low atmospheric temperatures.

Most recently, measurements of ion composition and temperature on the Ariel satellite (1962*o*1) provide additional evidence for the variation of thickness in the helium ion layer.[18] It is expected that these experimental data will allow further refinements of the model concept presented here.

* From the final analysis of a recent ion density profile (reference 5) it appears, that the equal-concentration levels at the corresponding temperature of 800°K are about 200 km lower than indicated in figure 1, while the thickness of the helium ion layer is in good agreement with the present model. This fact seems to imply an even stronger temperature-dependence of the ion concentration ratios η_{ij} than was assumed here.

REFERENCES

1. Nicolet, M., "Helium, an Important Constituent in the Lower Exosphere," *J. Geophys. Res.* **66**(7): 2263–2264, July 1961.
2. Hanson, W. B., "Upper-Atmosphere Helium Ions," *J. Geophys. Res.* **67**(1): 183–188, January 1962.
3. Bauer, S. J., and Jackson, J. E., "Rocket Measurement of the Electron-Density Distribution in the Topside Ionosphere," *J. Geophys. Res.* **67**(4): 1675–1677, April 1962.
4. Bourdeau, R. E., Whipple, E. C., Jr., Donley, J. L., and Bauer, S. J., "Experimental Evidence for the Presence of Helium Ions Based on Explorer VIII Satellite Data," *J. Geophys. Res.* **67**(2): 467–475, February 1962.
5. Donley, J. L., Private communication, 1962.
6. Taylor, H. A., Brinton, H., and Smith, C. R., "Instrumentation for Atmospheric Composition Measurements," in: *Proc. 8th Aero-Space Instrumentation Symposium, Washington, May 1962*, Pittsburgh: Instrument Society of America (In Press).
7. Hale, L. C.: "Ionospheric Measurements with a Multigrid Potential Analyzer," *J. Geophys. Res.* **66**(5): 1554, May 1961 (Abstract).
8. Pannekoek, A., "Ionization in Stellar Atmospheres," *Bull. Astronom. Inst. Netherlands* No. 19, July 7, 1922.
9. Eddington, A. S., "The Internal Constitution of the Stars," New York: Dover Publications, 1959, pp. 272–274.

10. DUNGEY, J. W., "Electrodynamics of the Outer Atmosphere," in: *The Physics of the Ionosphere,* London: Physical Society, 1955, pp. 229–236.
11. MANGE, P., "The Distribution of Minor Ions in Electrostatic Equilibrium in the High Atmosphere," *J. Geophys. Res.* **65**(11): 3833–3834, November 1960.
12. BAUER, S. J., "On the Structure of the Topside Ionosphere," *J. Atmos. Sci.* **19**(3): 276–278, May 1962.
13. BATES, D. R. and PATTERSON, T. N. L., "Hydrogen Atoms and Ions in the Thermosphere and Exosphere," *Planet. Space. Sci.* **5**(4): 257–273, August 1961.
14. NICOLET, M. and KOCH ARTS, G., "Helium and Hydrogen Escape," in: *Space Research III: Proc. 3rd Internat. Space Sci. Sympos., Washington, May 1962,* Amsterdam: North-Holland Publ. Co. (To be published).
15. HANSON, W. B. and ORTENBURGER, I. B., "The Coupling between the Protonosphere and the Normal F Region," *J. Geophys. Res.* **66**(5): 1425–1435 May 1961.
16. BOURDEAU, R. E. and BAUER, S. J., "Structure of the Upper Atmosphere Deduced from Charged Particle Measurements on Rockets and the Explorer VIII Satellite," in *Space Research III: Proc. 3rd Internat. Space Sci. Sympos., Washington, May 1962,* Amsterdam: North-Holland Publ. Co. (To be published).
17. ULWICK, J. C. and PFISTER, W., "Spatial and Temporal Variations of Electron Density from an Orbiting Satellite," in: *Space Research III: Proc. 3rd Internat. Space Sci. Sympos., Washington, May 1962,* Amsterdam: North-Holland Publ. Co. (To be published).
18. WILLMORE, A. P., Private communication, 1962.

THE ORIGIN OF THE ATMOSPHERES OF VENUS AND THE EARTH

A. G. W. CAMERON

Goddard Space Flight Center

Evidence regarding the Earth's atmosphere is reviewed and it is concluded that there are no detectable traces of the primitive solar nebula surviving in it. Evidence concerning the Venus atmosphere is analyzed and it is concluded that it probably contains of the order of 100 times the amount of nitrogen in the Earth's atmosphere and negligible amounts of water vapor. It is very difficult to see how the two atmospheres can have had a common origin, and it is proposed that the bulk of the Venus atmosphere is a remnant of the primitive solar nebula. In order to account for the absence of primitive gases from the Earth's atmosphere, a modified form of the Darwin–Wise hypothesis is proposed, according to which the Moon is formed as a result of the Earth's rotational instability induced by formation of the iron core. During such rotational instability the primitive gases would be entirely lost. A discussion is given of the circumstances surrounding the period of rotational instability.

INTRODUCTION

It has been evident for some time that the large fractionation between the rare gases and the common constituents in the atmosphere relative to solar composition implies that the bulk of the Earth's atmosphere has been outgassed from the Earth (Brown, 1952). If the Earth as a whole is of chondritic composition, then the amount of A^{40} in the atmosphere represents a degassing of about 10% of the K^{40} that would have decayed by now in the interior (Turekian, 1963). The main degassing appears to take place from volcanos, and the oceans appear to have been produced by the same degassing process as the atmosphere (Rubey, 1951).

The writer has recently examined the problems raised by the fact that the isotopic composition of atmospheric xenon is totally different from that of the meteoritic xenon which was probably entrapped from the primordial solar nebula from which the planets were formed (Cameron, 1962, 1963a). The atmosphere lacks the large excess of Xe^{129} that would exist if the Earth were formed at the same time as the meteorite parent bodies and were fully outgassed, and which would arise from the decay of I^{129}. The atmosphere contains much xenon of fission origin, probably resulting from the decay of Pu^{244}. The nonfissionogenic xenon in the atmosphere has been altered in isotopic composition in a pattern which the writer attributes to the effects of neutron capture in the Sun during the deuterium-burning phase of its original contraction. A quantitative examination of this point indicates that the majority of the nonfissionogenic xenon in the atmosphere would be required to have once been in the Sun and subsequently captured by the Earth (Cameron, 1963a). In a recent compilation of meteoritic abundances between which krypton and xenon can be interpolated, the writer has found that the atmospheric ratio of krypton to xenon is about twice the ratio that can be expected to result from nucleosynthesis (Cameron, 1963b). The most straightforward interpretation of this result is that the outgassing of the nonfissionogenic xenon from the interior of the Earth is much less efficient than that of krypton and presumably the other rare gases. In particular it should be noted that there is no evidence whatever that any gas from the primitive solar nebula was captured by the Earth and now survives in the atmosphere.

A straightforward attempt to predict the composition of the atmosphere of Venus would utilize the similarity of the mass and mean density of that planet to those of the Earth. One might expect that volcanos on Venus would outgas the same substances in essentially the same proportion as on the Earth. The water vapor in the Earth's atmosphere produces a greenhouse effect which raises the temperature at the surface of the Earth significantly above that which would exist without it. If the Earth were closer to the Sun the greenhouse effect would be substantially increased owing to the greatly enhanced evaporation from the oceans, and it is possible that this could cause the surface temperature to "run away" until most or all of the water from the oceans existed as steam in the atmosphere (Gold, 1963). On this picture the high surface temperature of Venus (600°K on the dark side, probably 700–800°K on the bright side) requires that the ground be at an optical depth of at least 60 in the infrared (Jastrow and Rasool, 1963), which should easily be provided by the 100 or more atmospheres of steam that would exist under these conditions.

THE ATMOSPHERE OF VENUS

The above straightforward speculation regarding the Venus atmosphere is not in accordance with various measurements.

The only constituent of the atmosphere of Venus which has been unambiguously determined by spectroscopic methods is CO_2, which is present to the extent of 4% by mass (Spinrad, 1962a). Spinrad (1962b) has carefully examined a high-dispersion spectrogram of Venus taken by Adams and Dunham at a time when there were evidently breaks in the clouds, and he has concluded that the mixing ratio of water vapor down to a pressure of 8 atm was less than one part in 10^5. This makes it unlikely that the clouds can be composed of water. Urey (1959) has noted that there cannot be more than 80 meter-atm of oxygen above the clouds, and CH_4, NH_3, and oxides of nitrogen cannot be important constituents of the Venus atmosphere. It is likely that the bulk of the atmosphere is composed of nitrogen.

Further information about the possible water vapor content of the atmosphere can be deduced from radar observations of Venus. Reflections from the solid surface of Venus have been observed at 12.5 cm and at 68 cm, and their characteristics have been analyzed by Muhleman (1963). At 12.5 cm a total of $11.2 \pm 2.8\%$ of the power that would be reflected from an equivalent conducting sphere in vacuum was actually returned, and at 68 cm approximately 10% of the power was returned under the same assumptions.

The microwave absorption coefficient in gases is experimentally observed and theoretically expected to vary as the square of the frequency. Hence the above radar results indicate that 12.5 cm waves are not attenuated by more than 10% in a double traversal of the Venus atmosphere. Steam is particularly effective in absorbing microwaves owing to the nonlinearity of its molecule. P. Thaddeus (private communication) has calculated that 3 gm/cm² of steam would produce a 10% attenuation of a 12.5-cm signal traversing a double thickness of an isothermal atmosphere with a pressure at the ground ~100 atm and a temperature of 675°K. Hence this becomes an upper limit to the amount of water vapor in the Venus atmosphere. It is consistent with Spinrad's determination of the upper limit to the water vapor mixing ratio. It is half the mean water vapor content of the Earth's atmosphere and immensely less than our naive expectation for the Venus atmosphere.

This leads us to consider how the high surface temperature of Venus can be maintained if there is no appreciable water vapor content in the atmosphere. Opik (1961) has suggested that the dry atmosphere is rendered opaque by clouds of dust and that the high surface temperature is created by wind friction. However, this model predicts that there should be no variation in the observed microwave brightness temperature of Venus as a function of its phase, which appears to contradict the recent measurements (Drake, 1963). It appears most probable that the high surface temperature is an extreme greenhouse effect (Sagan, 1960).

Jastrow and Rasool (1963) have shown that if the atmosphere of Venus is in radiative equilibrium, its infrared optical depth must be at least 60. Non-gray effects would tend to increase the amount of atmosphere required. The statement of

the optical depth does not specify the mass of the atmosphere: this requires additional relations between pressure, density, and opacity. If any part of the atmosphere is in convective equilibrium, then the temperature gradient will be lowered relative to an atmosphere in radiative equilibrium, and hence a specified ground temperature would only be reached at still higher values of the optical depth.

Plass and Stull (1963) have calculated the transmission through several thicknesses of CO_2 at various pressures: for an atmosphere of nitrogen and CO_2 with approximately Spinrad's mixing ratio they find that the optical depth becomes about 4 near 30 atm pressure. Evidently this is a much too small lower limit to the probable ground pressure on Venus. However, the opacity should increase rapidly with additional pressure.

P. Thaddeus (private communication) has obtained new laboratory measurements of the microwave absorption in N_2–CO_2 mixtures, and he has extrapolated the results in such a way that they can give an upper limit to the mass of the Venus atmosphere. His preliminary value for the microwave absorption coefficient with 5% CO_2 by volume is

$$\alpha = (1.4\times10^{-7})(\nu^2P^2/T^{3.7}), \quad (\text{cm}^{-1})$$

where ν is in cm^{-1}, P in atmospheres, and T in units of 100°K. If one assumes an isothermal atmosphere with a scale height $H=kT/mg$, then the optical depth for a double traversal of the atmosphere is

$$\tau = (1.4\times10^{-5})(k\nu^2P_0^2/mgT^{2.7}),$$

where P_0 is the ground pressure in atmospheres. With $\tau=0.1$, $\nu=0.080$ cm^{-1} (for the highest frequency radar observation), $T=675$°K, $g=877$ cm/sec^2, and $m=29/N_0=4.8\times10^{-23}$ gm, we have $P_0=240$ atm. Hence this is an upper limit to the Venus atmosphere imposed by the radar measurements.

Thus it appears that an atmosphere of N_2–CO_2 with a ground pressure of the order of 100 atm is consistent with the various measurements discussed above. Sagan (1962) has presented additional evidence which points toward a very massive atmosphere.

It is very difficult to see how this atmosphere can have an origin similar to that of the Earth. There is very little nitrogen in the surface rocks of the Earth, and if the Earth were fully outgassed one would not expect to get more than one order of magnitude greater amount of nitrogen into the atmosphere.

Accordingly, I wish to put forward the hypothesis that most of the Venus atmosphere was captured from the primitive solar nebula and has been retained throughout the subsequent history of the planet.

There are a number of problems that require discussion in connection with this hypothesis. The writer has published arguments in favor of an extremely massive solar nebula from which the planets are formed before the Sun and represent a very inefficient collection of condensable material (Cameron, 1962, 1963c). It has been very difficult to understand how the solar nebula could be dissipated without significant amounts of it being captured by the planets. According to this new hypothesis, such capture appears to have happened in the case of Venus and presumably also in the cases of all the other planets except the Earth. Thus to account for the differences between the atmospheres of Venus and the Earth it is necessary to find a mechanism whereby the Earth can have lost its primitive atmosphere. Such a mechanism is suggested in the following section.

In the primitive solar nebula carbon, nitrogen, and oxygen would have relative abundances of 16.6:3.0:29 (Cameron, 1963b). These elements would be captured into the primitive Venus atmosphere mostly in the form of CH_2, NH_2, and H_2O. As time goes on, photodissociation and a high exospheric temperature will lead to the escape of hydrogen, and the composition will be changed to N_2 and CO_2, with a slight excess of carbon. However, we note that we would have much more CO_2 than N_2, the reverse of the present situation. Thus it appears necessary to find a way to remove the great bulk of the CO_2 once in the Venus atmosphere.

On the Earth CO_2 is slowly removed from the atmosphere by interaction with the magnesium and calcium silicates (Urey, 1952). Magnesium and calcium carbonates and quartz are formed. The reactions require continued weathering of rock surfaces and the presence of liquid water as a

catalyst. Sagan (1962) has shown that there is much more CO_2 in the Venus atmosphere than would exist if the above reactions were in equilibrium. However, there is no liquid water on Venus and hence the reactions cannot proceed as they do on the Earth. Evidently it is necessary to enquire whether other catalysts exist or whether the reactions can proceed directly at the high temperatures characteristic of the surface of Venus. In this connection it should be noted that the early N_2–CO_2 atmosphere would be far more opaque than the present one, and consequently it is likely that the surface temperature of Venus was once much higher than at present.

The validity of the hypothesis also requires that whatever water is outgassed from Venus volcanos should be destroyed. Much of this destruction can probably be done by the excess carbon resulting from the reduction of the primitive atmospheric composition. Actually, much of the oxygen mentioned above in the solar nebula is tied up in the form of metal oxides and silicates. Taking into account the relative abundances of the metals and silicon in the solar nebula (Cameron, 1963b), the amounts of carbon, nitrogen, and oxygen left over for the atmospheric composition are approximately in the ratios 16.6:3.0:24. On this basis 4.6 parts of carbon are in excess of CO_2 and are available to form additional CO_2 with outgassed water. If the present amount of nitrogen in the Venus atmosphere approaches the upper limit derived above, then this excess carbon would be sufficient to combine with all the water in the terrestrial oceans.

It is also worth enquiring whether we should expect as much water to be outgassed on Venus as on the Earth. In the Earth the early radioactive heating first melted iron at a depth of a few hundred kilometers (MacDonald, 1959). In the case of Venus with its much higher ground temperature, iron would first be melted much nearer the surface. It seems likely that this would give additional opportunity for the iron to come in contact with water at depth in Venus, thus producing iron oxides with the evolution of hydrogen. MacDonald (1963) has pointed out that the extra plasticity of the surface rocks associated with the higher temperature has probably been responsible for the enhanced solar tidal drag that would be required to slow down the rotation Venus from a rate comparable to that of the other planets.

Many of these considerations regarding composition should also apply to the atmosphere of Mars, which also appears to be a N_2–CO_2 mixture. Thus the atmosphere of Mars could be derived from the solar nebula. However, the much smaller mass of the Martian atmosphere relative to the terrestrial atmosphere does not serve as an argument against outgassing as in the case for Venus. There is some indication that outgassing may be of much reduced importance on Mars relative to the Earth, since Mars has no detectable mountains (de Vaucouleurs, 1954).

LOSS OF THE PRIMITIVE TERRESTRIAL ATMOSPHERE

In seeking a mechanism by which the Earth can have lost its primitive atmosphere, we should seek to satisfy two principles:

(a) The process must have been complete; it is evident that the primitive atmosphere cannot have merged smoothly into the present outgassed atmosphere.

(b) The process must be one that either should not or clearly did not also operate in the case of Venus.

Apart from the fact that Venus is closer to the Sun than is the Earth, the most striking difference between the two planets is the fact that the Earth possesses a satellite. Moreover, it is a particularly anomalous satellite, since it is much more massive relative to its primary than is the case for any other satellite in the solar system, and its mean density is much less than those of the inner planets of the solar system or that of chondritic meteoritic material. Furthermore, there are serious problems connected with the previous dynamical history of the Earth-Moon system.

Munk and MacDonald (1960) have found that dissipative processes are removing kinetic energy from the Earth-Moon system at about three times the rate previously obtained by Jeffreys (1959). Tidal friction in the seas appears inadequate to account for this, and it seems likely that the bulk of the dissipation occurs within the solid Earth. This renders it unlikely that the

efficiency of the dissipative mechanism can have changed rapidly with time, and extrapolations of the dynamical motions toward the past can be made with some degree of confidence. MacDonald (1961) has found that the Moon would recede from 10^5 km to its present distance in only $1.3\times 10_2$ years, and this time would be considerably reduced if nonlinear interactions were taken into account. There appear to be some conditions under which the time for recession of the Moon can be significantly lengthened (Slichter, 1963), but it seems extremely difficult to lengthen the time to the full age of the Earth, 4.5×10^9 years.

There has been much speculation regarding the origin of the Moon. Darwin (1898) postulated that the Moon was once part of the Earth, but that when the solar tide came into resonance with the free oscillations of the Earth, enormous tidal bulges were raised, one of which broke off to become the Moon. This idea was abandoned when it was realized that friction would severely limit the height of any such tidal bulge (Jeffreys, 1930). Since that time most authors have variously thought that the Moon was formed independently of the Earth, and perhaps was captured by it, despite all the difficulties in any such capture hypothesis.

The idea that the Moon could have arisen from rotational instability of the Earth has recently been revived by Wise (1963), who pointed out that the formation of the Earth's core would have reduced the moment of inertia, and thus perhaps the threshold for rotational instability could be exceeded.

The basic relevance of these ideas to the general subject of the present paper is the following: If the Earth was ever rotationally unstable at the equator then any primitive atmosphere would be entirely lost. Hence the rotational instability hypothesis satisfies our two criteria; it provides a complete loss of the primitive atmosphere, and the process has not occurred for Venus if the lack of a satellite revolving about that planet can be taken as an indication. We shall therefore examine the rotational instability hypothesis in more detail.

Wise (1963) assumed that as the spin of the Earth increased, the figure deformed from an oblate Maclaurin spheroid to a cigar-shaped Jacobian triaxial ellipsoid. After this had reached limiting axial ratios of 8:10:23 it would deform into a pear-shaped Poincare figure spinning with the long axis in the plane of rotation. However, this great elongation of the Earth would appear to involve a progressive increase in its moment of inertia, rather than the decrease which would be provided by the formation of the iron core. Furthermore, Elsasser (1963) has argued that the formation of the Earth's core must have been a very catastrophic process, since the viscosity of silicate materials should greatly increase with increasing pressure towards the interior. Since iron must first melt near the surface of the Earth, large pools of it must collect before the silicate materials will be fractured, thus allowing the pools to spill towards the center. If this is correct, then rotational instability would suddenly set in at the equator, and there would be no opportunity for the slow plastic deformation discussed by Wise.

Let us make some simple calculations designed to be illustrative of the problem rather than precise. Assume that the masses of the Earth and the Moon are combined in one body with conservation of total volume and uniform density. The radius of this body would be $R=6.41\times10^8$ cm. If the total angular momentum now residing in the Earth-Moon system were given to this body, it would spin with an angular velocity of 3.47×10^{-4} radians/sec. For comparison, the Kepler orbital angular velocity at which the equator would become rotationally unstable is 1.23×10^{-3} radians/sec. However, let us put aside for the moment the discrepancy in these two numbers and consider our combined body to be rotating at the verge of rotational instability.

Now suppose a core to form of radius 3473 km and mean density (assumed uniform) of 11.3 gm/cm^3. Conservation of angular momentum requires

$$I_f\omega_f=I_i\omega_i,$$

where the subscripts f and i refer to final and initial conditions respectively, I is a moment of inertia, and ω is an angular velocity. With the numbers given above we find $\omega_f=1.45\times10^{-3}$ radians/sec, which significantly exceeds the limit of rotational instability.

As stated above, it seems likely that the

formation of the core will be an extremely catastrophic process which will lead to extensive shattering throughout the Earth and particularly at the surface. Thus we may assume that rotational instability at the equator will lead to the departure of large numbers of small bodies from the Earth. The Earth will reform a spherical shape behind them. Suppose that the net effect is to remove a surface layer of thickness a and mass m and let it revolve at a distance R equal to the original equatorial radius. The remainder of the Earth will be left on the verge of rotational instability, which for convenience we shall assume to be the initial angular velocity ω_i. The mass loss is

$$m=\tfrac{4}{3}\pi\rho_m[R^3-(R-a)^3],$$

where ρ_m is the density of the mantle from which the core material has been removed. The equation for conservation of angular momentum is

$$I_c\omega_i+\tfrac{2}{5}(M_m-m)\frac{[(R-a)^5-R_c^5]}{[(R-a)^3-R_c^3]}\omega_i+mR^2\omega_i=I_i\omega_i$$

where I_c and R_c are the moment of inertia and radius of the core and M_m is the mass of the total mantle after removal of the core material.

The above two equations must be solved for m and a. The results are

$$a=3.59\times10^7 \text{ cm},$$

$$m=7.58\times10^{26} \text{ gm}.$$

This mass loss is 10.3 times the mass of the Moon.

The Moon cannot immediately be accumulated from the pieces of matter thrown off at the equator since they are all well inside the Roche limit at 2.89 Earth radii. The dimensions of the ejected bodies will depend on the strength of their materials and upon the amount of shattering that accompanies the violent events at the time of core formation. If dimensions of as much as some tens of kilometers are possible, then the fragments may be able to raise sufficient tides in the body of the Earth to be dispersed to larger distances at which accumulation is possible. Collisions among the bodies can in principle also result in a dispersion toward larger distances, but at the expense of returning the majority of the material to the Earth, in order to conserve the total angular momentum.

Let us determine a limit on the mass that could be placed at the Roche limit, $R_r=2.89\ R$, owing to such a scattering process. We replace $mR^2\omega_i$ in the angular momentum equation by $mR_r^2\omega_r$, where ω_r is the Kepler angular velocity at the Roche distance (8.05×10^{-4} radians/sec). We then obtain

$$a=3.49\times10^6 \text{ cm},$$

$$m=7.73\times10^{25} \text{ gm}.$$

This mass loss is 1.05 times the mass of the Moon.

It is evident that, while the scattering process may make an important contribution to the amount of matter outside the Roche limit, the above calculation does not give enough margin for the inefficiencies of the problem, particularly the intrinsic spread in radial distances that would be involved and the unlikely probability of having the initial angular velocity of the Earth so close to the value for rotational instability. Hence it is likely that the tidal effects of the individual fragments must play an important role in causing the recession of much of the mass to the Roche limit if the Moon has originated by this general mechanism.

Let us now assume that the tidal recession of the fragments has been the only operating process and calculate the minimum initial angular momentum needed to throw off a lunar mass at the surface of the Earth. We return to the initial angular momentum equation and replace the term $I_i\omega_i$ by $I_i\omega_0$, where ω_0 is the desired angular velocity. Then, with $m=7.35\times10^{25}$ gm and $a=3.33\times10^6$ cm, we obtain

$$\omega_0=1.05\times10^{-13} \text{ radians/sec}.$$

This is 86% of rotational instability.

One additional point that favors a considerable tidal recession of the individual fragments concerns a possible explanation of the discrepancy between the present angular momentum of the Earth-Moon system and that required to produce rotational instability. There will have been some slowing of the Earth's rotation due to solar tides, but these are so much less than the present lunar tides that it appears unlikely that they can account for this discrepancy. However, it should be noted that perturbations of the motion of a small third body in a two-body system can easily

result in the ejection of the third body from the system. Hence the accumulation of the Moon is likely to have been a rather inefficient process, with the ejection from the Earth-Moon system of a large number of the small bodies which make close approaches to the growing Moon. Even the bodies which collide with the growing Moon are unlikely to be fully retained if the collision involves an appreciable relative velocity, since many of the fragments formed in such collisions will have more than the escape velocity. The loss of such material from the Earth-Moon system will be accompanied by a considerable decrease in the angular momentum of the system.

We can easily place limits on the amount of mass loss that would be required to reduce the angular momentum to the present value. Let us suppose that there is some principal accumulation distance R_a at which a total mass m is transferred from a circular to a parabolic orbit. The total angular momentum carried away by this mass is $2^{1/2}mR_a^2\omega_a$, where $\omega_a=(GM/R_a^3)^{1/2}$ is the Kepler angular velocity at R_a and M is the mass of the Earth. This may be set equal to the angular momentum deficit of the system, $I_i(\omega_i-\omega_j)$, where $I_i\omega_j$ is the angular momentum of the homogeneous (Earth+Moon) on the verge of rotational instability, and $\omega_j=3.47\times10^{-4}$ radians/sec is the angular velocity which, as we have seen, the homogeneous (Earth+Moon) would have if the present angular momentum of the system were given to it. Hence we obtain

$$mR_a^{1/2}=I_i(\omega_i-\omega_j/(2GM)^{1/2}$$
$$=3.12\times10^{31}.$$

If $R_a=R_r$, the Roche limit, then

$$m=7.28\times10^{26}\text{ gm}$$
$$=9.90\text{ times the mass of the Moon.}$$

If $R_a=3.84\times10^{10}$ cm, the present lunar distance,

$$m=1.59\times10^{26}\text{ gm}$$
$$=2.17\text{ times the mass of the Moon.}$$

Since the accumulation is much more likely to have taken place near the Roche distance than near the present lunar distance, the present proposals require that tidal recession should be more important than scattering in removing the mass from the immediate vicinity of the Earth following the core formation catastrophe.

Thus we have arrived at a picture in which the amount of matter lost by the Earth-Moon system is considerably larger than the mass of the Moon and very greatly larger than the total mass of the asteroid belt. Owing to the fact that this material will continue to cross the orbital path of the Earth, it cannot have a stability comparable to that of the asteroids. In a time comparable to the cosmic ray exposure ages of the meteorites, the majority of the material will either collide with the Earth or be perturbed sufficiently in its motion so that its orbital path will cross that of Jupiter, thus leading either to collision with that giant planet or to ejection from the solar system. It is possible that some of the asteroidal bodies crossing the orbits of the inner planets are remnants of the material ejected from the Earth-Moon system.

Because of the extremely disruptive character of the events postulated above, it is unlikely that any surface features of the Earth would survive in recognizable form. Thus it is unlikely the events can have occurred less than three billion years ago, the age of the oldest rocks. Elsasser (1963) has argued that the crust of the Earth would form more slowly than the core. MacDonald (private communication) has suggested that the principal tidal friction occurs at the boundaries between the continental blocks. Thus it seems that tidal friction would be diminished relative to the present rate after formation of the Moon, so that it may be possible to reconcile the present dynamics of the Earth-Moon system with an age of three billion years.

It should be recalled that iron will first melt in the Earth at a depth of several hundred kilometers (MacDonald, 1959). If the Moon were formed from the material at the surface of the Earth, it would presumably not represent differentiated material. Urey (1960) has pointed out that the low density of the Moon requires an abnormal composition; a possible composition would be chondritic material which had lost about half of its iron. It seems likely that such a composition would be attained as a result of the very deep disruption of the Earth discussed above, taken together with a random selection of the material assembled to make the Moon.

From experimental aerodynamic evidence Chapman and Larson (1963) have argued that many tektites, particularly australites, must have come from the Moon. The chemical composition of the tektites then demands that they be produced from highly differentiated rocks such as granites. O'Keefe and W. S. Cameron (1962) have suggested mechanisms for producing granites on the Moon. It may be that the preliminary differentiation of rocks discussed in the present paper and assumed to precede formation of the Moon would assist in the subsequent differentiation discussed by O'Keefe and Cameron.

DISCUSSION

The evidence regarding the differences between the compositions of the atmospheres of Venus and the Earth discussed in this paper suggests that a planet will ordinarily acquire its atmosphere from the primitive solar nebula. However, the Earth is an exception to this rule, and hence it seems necessary to link this new cosmogonic hypothesis to the old cosmogonic hypothesis of Darwin regarding the origin of the Moon which Wise has revived in new form. Upon developing Darwin's hypothesis in a manner somewhat different from that of Wise, we see that it is apparently possible to reconcile many puzzling features of the Moon and the Earth-Moon system.

It is evident that this new set of cosmogonic hypotheses has profound philosophical implications, particularly when we try to extrapolate conditions on the Earth to determine conditions on other planets. It has been found that highly complex molecules are formed in "primitive" atmospheres of both the kind that would be accreted from the solar nebula and that would be outgassed from the Earth (Miller, 1953; Abelson, 1956). Probably life could originate in either atmosphere, but it would probably develop very differently. We do not know how likely a planet is to be formed with the possibility of reaching rotational instability and developing a massive satellite. If this is a very rare occurrence, then the arguments of similarity usually used in an attempt to deduce the character of life near other stars (Cameron, 1963d) need revision.

ACKNOWLEDGMENTS

I am particularly indebted to Dr. Patrick Thaddeus for allowing me to use his experimental and theoretical results on microwave absorption before publication. Various aspects of the above proposals grew out of stimulating conversations with F. L. Whipple, G. J. F. MacDonald, J. A. O'Keefe, H. H. Hess, J. Geiss, and E. Anders. I am grateful to N. Smith and S. P. Maran at the University of Michigan for pointing out an error in a preliminary version of this paper.

REFERENCES

Abelson, P. H. Amino-acids formed in "primitive atmospheres." *Science* **124,** 935, 1956.

Brown, H. Rare gases and the formation of the earth's atmosphere. *In* "The Atmospheres of the Earth and Planets" (G. P. Kuiper, ed.). Univ. of Chicago Press, Chicago, Illinois, 1952.

Cameron, A. G. W. The formation of the Sun and planets. *Icarus* **1,** 13, 1962.

Cameron, A. G. W. The origin of atmospheric xenon. *Icarus* **1,** 314, 1963a.

Cameron, A. G. W. "Nuclear Astrophysics." Notes from lectures given at Yale University, 1963b.

Cameron, A. G. W. Formation of the solar nebula. *Icarus* **1,** 339, 1963c.

Cameron, A. G. W. (ed.). "Interstellar Communication." W. A. Benjamin, New York, 1963d.

Chapman, D. R., and Larson, H. K. On the lunar origin of tektites. *J. Geophys. Res.* **68,** 4305, 1963.

Darwin, G. H. "The Tides." Houghton-Mifflin, Boston and New York, 1898.

de Vaucouleurs, G. "Physics of the Planet Mars." Faber and Faber, London, 1954.

Drake, F. D. Ground-based microwave observations of Venus in 1962. *Trans. Am. Geophys. Union* **44,** 71, 1963.

Elsasser, W. M. Early history of the earth. *In* "Earth Science and Meteorites" (J. Geiss, and E. D. Goldberg, eds.). North-Holland Publ. Co., Amsterdam. 1963.

Gold, T. Comments on outgassing. *In* "Origin and Evolution and Oceans" (A. G. W. Cameron, and P. Brancazio, eds.). In preparation, 1963.

Jastrow, R., and Rasool, S. I. Radiative transfer in the atmospheres of Venus and Mars. *In* "Space Research III" (W. Priester, ed.). North-Holland Publ. Co., Amsterdam, 1963.

Jeffreys, H. The resonance theory of the origin of the moon, II. *Monthly Notices Roy. Astron. Soc.* **91,** 169, 1930.

Jeffreys, H. "The Earth." Cambridge Univ. Press, London, 1959.

MacDonald, G. J. F. Calculations on the thermal history of the earth. *J. Geophys. Res.* **64,** 1967, 1959.

Macdonald, G. J. F. Interior of the moon. *Science* **133,** 1045, 1961.

MacDonald, G. J. F. The internal constitutions of the inner planets and the moon. *Space Sci. Rev.*, in press, 1963.

Miller, S. L. A production of amino acids under possible primitive earth conditions. *Science* **117,** 528, 1953.

Muhleman, D. O. The electrical characteristics of the atmosphere and surface of Venus from radar observations. *Icarus* **1,** 401, 1963.

Munk, W. H., and MacDonald, G. J. F. "The Rotation of the earth." Cambridge Univ. Press, London, 1960.

O'Keefe, J. A., and Cameron, W. S. Evidence from the Moon's surface features for the production of lunar granites. *Icarus* **1,** 271, 1962.

Opik, E. J. The acolosphere and atmosphere of Venus. *J. Geophys. Res.* **66,** 2807, 1961.

Plass, G. N., and Stull, V. R. Carbon dioxide absorption for path lengths applicable to the atmosphere of Venus. *J. Geophys. Res.* **68,** 1355, 1963.

Rubey, W. W. Geologic history of sea water. *Bull. Geol. Soc. Am.* **62,** 1111, 1951.

Sagan, C. The radiation balance of Venus. *Jet Propulsion Lab., Calif. Inst. Technol., Tech. Rept.* **32–34,** 1960.

Sagan, C. Structure of the lower atmosphere of Venus. *Icarus* **1,** 151, 1962.

Slichter, L. B. Secular effects of tidal friction upon the earth's rotation. *J. Geophys. Res.* **68,** 4281, 1963.

Spinrad, H. Spectroscopic temperature and pressure measurements in the Venus atmosphere. *Publs. Astron. Soc. Pacific* **74,** 187, 1962a.

Spinrad, H. A search for water vapor and trace constituents in the venus atmosphere. *Icarus* **1,** 266, 1962b.

Turekian, K. K. Degassing of argon and helium from the earth. *In* "Origin and Evolution of Atmospheres and Oceans" (A. G. W. Cameron and P. Brancazio, eds.). In preparation, 1963.

Urey, H. C. "The Planets: Their Origin and Development." Yale Univ. Press, New Haven, Connecticut.

Urey, H. C. The atmospheres of the planets. *Handbuch der Physik* **52,** 363, 1959.

Urey, H. C. Lines of evidence in regard to the composition of the moon. *In* "Space Research" (H. K. Bijl, ed.). North-Holland Publ. Co., Amsterdam.

Wise, D. U. An origin of the moon by rotational fission during formation of the earth's core. *J. Geophys. Res.* **68,** 1547, 1963.

Note added in proof.—It has been drawn to my attention that a possible connection between rotational instability of the Earth and loss of a primitive atmosphere, together with formation of the Moon, was earlier suggested by A. E. Ringwood [*Geochim. Cosmochim. Acta* **20,** 241 (1960)].

THE ORIGIN OF ATMOSPHERIC XENON

A. G. W. CAMERON

Goddard Space Flight Center

Some of the differences in the isotopic composition of atmospheric and meteoritic xenon can be attributed to irradiation of solar xenon by neutrons during the deuterium burning phase of solar contraction, which was followed by accretion of the irradiated xenon into the atmosphere. This problem is examined here more carefully in view of a new measurement which has been made of the reaction rate for the capture of protons by deuterium. It is concluded that, apart from the xenon produced by the decay of extinct radioactivities, most of the atmospheric xenon was in the sun during the deuterium burning phase.

INTRODUCTION

In a recent article by the author the differences between the isotopic composition of atmospheric and meteoritic xenon were analyzed.[1] The differences were attributed to three causes: (1) the Xe^{129} decay product of the extinct radioactivity I^{129}; (2) the fission products of the extinct radioactivity Pu^{244}; and (3) the composition changes produced by neutron capture in one of the sources of atmospheric xenon. A crude quantitative analysis of the third cause led to the conclusion that the neutron capture had occurred during the deuterium burning phase of solar contraction, and that about $\frac{1}{3}$ of the atmospheric xenon had once resided in the post-deuterium-burning sun.

Recently, Griffiths, Lal, and Scorfe (private communication from G. M. Griffiths to H. Reeves) have redetermined the reaction rate at low energies for the $D^2(p, \gamma)He^3$ reaction. This competes with the $D^2(d, n)He^3$ and $D^2(d, p)H^3$ reactions in the deuterium burning process. They have found that the $D^2(p, \gamma)He^3$ reaction rate is greater than that given by Salpeter[2] by a factor 2.4. Dr. Griffiths states that the error in the new value may be as large as 50 percent, but is probably somewhat smaller.

In view of this change, the question of the solar origin of some atmospheric xenon has been examined more carefully, and this paper reports the results.

NEUTRON PRODUCTION IN THE SUN

Equation 1 of reference 1 gives the ratio of probabilities for the reaction of a deuteron with a deuteron relative to that of a deuteron with a proton. Let us divide this ratio by the factor 2 in order to count only those deuteron-deuteron reactions that produce neutrons, and also let us utilize the new reaction rate of Griffiths, Lal, and Scorfe. The modified ratio then becomes:

$$\frac{P_n}{P_p} = 1.24 \times 10^5 \left(\frac{x_D}{x_H}\right) \exp\left(-\frac{5.4}{T_6^{\frac{1}{3}}}\right), \qquad (1)$$

where P_n and P_p are the reaction rates for producing neutrons and absorbing protons, x_D and x_H are the concentrations by weight of deuterium and hydrogen, and T_6 is the temperature in units of 10^6 °K ($T_6 = T \times 10^6$ °K).

It has been pointed out recently by Hayashi that the contracting sun must have been fully convective until it approached quite near the main sequence configuration.[3] This has been confirmed by Ezer and Cameron, who found that deuterium burning took place during the fully convective phase and lasted only about 3×10^5 years because of the relatively high luminosity of the sun at the time.[4] Thus we can be assured that all the material in the sun was subjected to the neutron flux associated with deuterium burning,

and that this occurred early in the history of the solar system.

With the previous assumptions of Cameron that $x_D/x_H = 3\times10^{-4}$ initially and that $T_6 = 0.8$,[1] it follows from equation 1 of this paper that $P_n/P_p = 0.111$. Hence, allowing for the fact that this is twice the average rate of neutron production, we find that there will be 0.046 neutron produced per initial deuteron present in the sun.

NEUTRON CAPTURE IN THE SUN

The neutrons produced during deuterium burning become thermalized and then are captured by surrounding nuclei in proportion to the capture cross sections. We are particularly interested in finding the number of neutrons captured by I^{127} to form, after beta decay, an excess abundance of Xe^{128}.

The capture cross section in I^{127} must be computed by summing the contributions from all effective resonances of the compound nucleus. Each resonance makes a contribution to the average cross section of

$$\langle\sigma(E_r)\rangle = 555\,\frac{g\,\Gamma_n\,\Gamma_\gamma}{T_6\,\Gamma}\exp\left(\frac{-0.01161\,E_r}{T_6}\right), \quad (2)$$

where E_r is the resonance energy in ev; $g = (2J+1)/2(2I+1)$ is a statistical factor, J and I being the compound and initial spins of the nucleus under consideration; and Γ_n, Γ_γ, and Γ are the neutron, radiation, and total widths of the resonance in ev. The most complete data for the resonances in the compound nucleus I^{128} appear to be those of Desjardins, et al.,[5] as quoted in the Nuclear Data Tables. The average cross section was computed from these data with the assumptions $g = 0.5$ and $\Gamma_\gamma = 0.1$ if not directly measured. The result is $\langle\sigma\rangle = 25$ barns. This is probably correct to about ten percent. Most of the neutrons produced will react with hydrogen and He^3. The capture in any other nucleus can be computed as a small perturbation on the neutron elimination by these two nuclei. The situation is slightly complicated by the fact that He^3 is produced by the deuterium burning process.

Both H^1 and He^3 have neutron elimination cross sections varying as $1/v$, where v is the neutron velocity. At $T_6 = 0.8$, the cross section averaged over the Maxwell distribution of velocities is equal to the actual cross section at 87.6 ev. Hence for H^1, $\langle\sigma\rangle = 0.0056$ barns, and for He^3, $\langle\sigma\rangle = 91$ barns. Neutron capture in H^1 produces additional deuterium, but we shall neglect the second order production of additional neutrons arising from this fact. The neutron interaction with He^3 gives the reaction $He^3(n, p)H^3$, but the tritium decays back to He^3 rapidly, compared to the deuterium burning time of the contracting sun. Hence we can neglect any depletion of He^3 due to neutron capture.

There appears to be remarkably little variation in the ratio He^3/He^4 in primordial gases extracted from meteorites (private communication from P. Signer). The average number ratio of these nuclei is approximately 3×10^{-4}. If we assume a number ratio of these nuclei is approximately 3×10^{-4}. If we assume a number ratio of He^4/H^1 equal to 0.156, then we obtain a primordial number ratio of He^3/H^1 equal to 4.7×10^{-5}. Multiplying this ratio by 91/0.0056, the ratio of neutron cross sections, we see that the primordial He^3 is 0.76 times as effective as H^1 for eliminating neutrons. During deuterium burning 0.954 of the deuterium was converted to He^3, and hence, with an assumed number ratio D^2/H^1 equal to 1.5×10^{-4}, the He^3 produced is 2.32 times as effective as H^1 in eliminating neutrons.

The number of neutrons produced per hydrogen atom is $0.046\times1.5\times10^{-4} = 6.9\times10^{-6}$. Hence the number of neutrons captured by any other nucleus with average capture cross section $\langle\sigma\rangle$ is approximately

$$f = \frac{6.9\times10^{-6}\,\langle\sigma\rangle}{0.0056}\int_0^1 \frac{dx}{1+0.76+2.32\,x} = 4.5\times10^{-4}\,\langle\sigma\rangle \quad (3)$$

In the case of particular interest here, I^{127}, $f=0.0113$.

DISCUSSION

In reference 1 the conclusions regarding the solar origin of some atmospheric xenon were based on table 3 of that paper, in which it was found that, relative to Xe^{124} and X^{126}, which should be only negligibly depleted by neutron capture, Xe^{128} was overabundant in the atmosphere by 15 ± 2 percent. This is to be attributed to neutron capture in I^{127} in the sun.

In the preceding section, we have found that about 0.0113 of the I^{127} should be destroyed by neutron capture in the sun. Relative to the normalization used in table 3 of reference 1, the abundance of I^{127} is approximately 8. Hence the excess Xe^{128} that should be formed during deuterium burning in solar material is 8×0.0113, or 9 percent.

A comparison of these two numbers seems to indicate that atmospheric xenon contains more excess Xe^{128} than can be accounted for by neutron capture in pure solar material. However, if we take the full 50 percent error allowable in the new determination of the $D^2(p, \gamma)He^3$ reaction rate, then it appears that the two numbers barely agree within the overlap of their errors.

Thus, this interpretation of the xenon isotopic anomalies leads to the conclusion that the majority of all xenon in the atmosphere that does not result from the decay of extinct radioactivities was once part of the sun during the deuterium burning phase. There is no evidence for any xenon surviving in the atmosphere from a primitive protoplanet in which the earth may be thought to have formed; such residual xenon could be at most a small fraction of the atmospheric xenon. Hence the majority of the noble gases in the atmosphere, except for helium and A^{40}, have probably accreted from the solar wind.[1] Although it was argued in reference 1 that the data of Krummenacher, Merrihue, et al.[6] on the isotopic composition of krypton were slightly inconsistent with this conclusion, the uncertainty in the solar abundance ratio of bromine and krypton renders the argument inconclusive. Further work krypton would be desirable.

If the noble gases can accrete from the solar wind, what about hydrogen? Could the deuterium to hydrogen ratio in terrestrial waters have been changed relative to the primordial ratio? Boato measured this ratio for several carbonaceous chondrites, and found D^2/H^1 ratios as high as 2×10^{-4}.[7] Thus it is quite possible that the primordial sun may have had considerably more deuterium than has so far been assumed in this paper.

If the calculations of this paper are repeated with $x_D=x_H=4\times10^{-4}$, then we obtain $f=0.0125$ for I^{127}. Hence on the normalization of table 3 of reference 1 we would expect a 10 percent excess Xe^{128} abundance in solar material. This gives a slight improvement in the agreement of the predicted and observed values. Clearly it is desirable to investigate the deuterium content of carbonaceous chondrites more thoroughly. The major source of error in the predicted value remains the $D^2(p, \gamma)He^3$ reaction rate, and it is particularly desirable to improve that measurement further.

REFERENCES

1. Cameron, A. G. W., "The Formation of the Sun and Planets," *Icarus* **1**(1): 13–69, May 1962.
2. Salpeter, E. E., "Nuclear Reactions in Stars. II. Protons on Light Nuclei," *Phys. Rev.* **97**(5): 1237–1244, March 1, 1955.
3. Hayashi, C., "Stellar Evolution in Early Phases of Gravitational Contraction," *Publ. Astronom. Soc. Japan* **13**(4): 450–452, 1961.
4. Ezer, D., and Cameron, A. G. W., papers presented at the 111th meeting of the Amer. Astronom. Soc., Yale Univ., New Haven, Conn., August 1962: "High-Luminosity, Fully Convective Contracting Sun. I," *Astronom. J.* **67**(9): 575, November 1962 (abstract); "High-Luminosity, Fully Convective Contracting Sun. II," *Astronom. J.* **67**(9): 572, November 1962 (abstract).
5. Desjardins, S., Havens, W. W., Jr., et al., "Neutron Resonances in Th, I, Br," *Bull. Am. Phys. Soc.* Ser. 2, **5**(4): 295, April 25, 1960 (abstract).
6. Krummenacher, D., Merrihue, C. M., et al., preprint, 1961.
7. Boato, G., "The Primeval Cosmic Abundance of Deuterium," *Phys. Rev.* **93**(3): 640–641, February 1, 1954.

THE INFRARED HORIZON OF THE PLANET EARTH

R. A. HANEL, W. R. BANDEEN AND B. J. CONRATH

Goddard Space Flight Center

Horizon sensors are essential elements in the orientation systems of many space vehicles and satellites. Their design and an evaluation of their capabilities and limitations requires a knowledge of the spectral radiance emitted by the Earth. This report investigates the discontinuity between the surface, the atmosphere, and outer space for various latitudinal, seasonal, climatic, and meteorological conditions. Calculations are carried out for the following five spectral intervals—the water vapor band from 6.33 to 6.85 microns, the ozone band from 8.9 to 10.1 microns, the atmospheric window from 10.75 to 11.75 microns, the carbon dioxide band from 14 to 16 microns, and finally the rotational water vapor band from 21 to 125 microns. The superior properties of the carbon dioxide band and the rotational water vapor region for horizon sensor purposes are shown.

INTRODUCTION

A knowledge of the direction toward the center of the Earth is required for the proper orientation of many scientific satellites and manned space capsules. Moreover, the required accuracy in determining the direction of the local vertical is constantly increasing with the growing sophistication and complexity of experiments.

Normally, a horizon sensor system determines the center of the apparent disk of the Earth, which is very close to the center of gravity. In order to make the system independent of sunlight, it has become customary to restrict the spectral response of the horizon sensor to the infrared part of the spectrum or, more specifically, to wavelengths longer than 2 microns. For the sake of simplicity and engineering convenience the spectral response has seldom been extended much beyond 15 microns (e.g., Wormser and Arck, 1962). This utilizes about one-half the energy of a 270°K blackbody, which is very important for normally energy-limited systems.

This approach is entirely justified as long as one is satisfied with an uncertainty in the horizon on the order of 50 km (about 1 deg for a 600-km orbit) and as long as the system is able to accommodate the changes in signal strength caused by varying meteorological conditions (Conrath, 1962). Only recently have horizon sensors been considered which are more selective spectrally.[1] However, for the design of advanced stabilization systems, a better understanding of the infrared structure of the horizon is required.

This report investigates the discontinuity between the Earth, the atmosphere, and outer space. The apparent horizon was computed for several wavelength regions and under a variety of atmospheric conditions. Latitudinal, seasonal, and climatic effects were considered, as well as the effect of clouds for several altitudes. Calculations are given in detail to benefit the reader who is interested in the problem of horizon sensing but who may not be familiar with the problems in the radiative transfer of energy.

While writing this paper it came to our attention that a similar investigation, although for somewhat different spectral intervals and model atmospheres, had been carried out recently by K. Y. Kondratiev and K. E. Yakushevaskaya (1962) at the University of Leningrad, U.S.S.R. For ease in comparing the two papers, we have plotted our results in Figs. 10 through 14 in a format similar to that used by Kondratiev and Yakushevskaya.

OUTLINE OF PROBLEM

In the spectral range of infrared emission, the sharpness of the horizon depends on the vertical

TABLE 1.—*Wavelength intervals considered*

Wavelength in microns	Major absorber	Secondary absorbers
6.33–6.85	H_2O	None
8.9–10.1	O_3	H_2O continuum
10.75–11.75	H_2O continuum	Dust
14–16	CO_2	O_3, H_2O
21–125	H_2O	None

temperature and pressure profiles and on the distribution of absorbing (and therefore emitting) matter in the atmosphere. Water vapor, carbon dioxide, and ozone have strong absorption bands in this range, which must be considered. In addition, droplets and ice crystals in clouds, and perhaps dust and aerosols must also be included.

We have made an attempt to compute the fraction of the radiance N ($w\ m^{-2}$ $ster^{-1}$) of the Earth and the atmosphere within particular wavelength regions where only one of the atmospheric constituents dominates absorption. Less important constituents have been considered only qualitatively, with the exception of the 8.9- to 10.1-micron region, where emission by both ozone and the water vapor continuum is taken into account. The wavelength intervals considered are listed in table 1.

The intervals between the quoted wavelength regions, such as between 12–14 microns and 16–21 microns, are characterized by overlapping bands and semi-transparent gaps. The radiance of these intervals is somewhere between that of the strong absorption regions of the gases discussed and that of the window chosen for our investigation. The ranges less than 6 microns and greater than 125 microns are of minor interest since the fraction of thermal energy available in these spectral ranges is small.

The fraction of the radiance emitted by the Earth and the atmosphere, which can be detected by a horizon sensor with a spectral response ϕ_λ, is given by

$$N_\phi = -\frac{1}{\pi}\int_{\lambda=0}^{\infty}\int_{u=0}^{\infty}\phi_\lambda B_\lambda(T)\frac{\partial\tau(u^*)}{\partial u^*}\,du^*\,d\lambda. \quad (1)$$

(For a discussion of radiative transfer equations see, e.g., Möller, 1951). In our calculation, the spectral response ϕ_λ of a possible detection system has been taken as unity within the quoted wavelength ranges and as zero outside. For this specific form of ϕ_λ we redefine $N_\phi = N_{\Delta\lambda}$ in (1). The integration of $N_{\Delta\lambda}$ was done graphically with the aid of radiation charts. The problem can be divided into the following tasks:

1. The formulation of an expression for the reduced optical path u^*, for a curved atmosphere as a function of height, and for various gas pressures, temperatures and concentrations;
2. The section of atmospheres of extreme but realistic properties to show the limits which can reasonably be expected in the apparent horizon;
3. Numerical computation of u^* for each gas and atmosphere selected;
4. Construction of a radiation chart for each spectral range under consideration;
5. The plotting and graphical evaluation of the radiation charts for all spectral ranges and atmospheres.

The results of the integrations are summarized in diagrams and the problem is discussed below in the sequence previously listed.

THE REDUCED OPTICAL PATH

The reduced optical path du^* of an absorber has been taken from Elsasser (1960), and its relation to the true optical path du is expressed by

$$du^* = (P/P_0)\,(T_0/T)^{1/2}\,du. \quad (2)$$

The factors P/P_0 and $(T_0/T)^{1/2}$ take into account the pressure and temperature dependence of line width. In our computations this expression for u^* was used for all gases, except for ozone when the pressure exceeded 10 cm Hg, after which a constant factor P_1/P_0 replaced P/P_0. As pointed out by Elsasser (1960), the results obtained by this replacement are in better agreement with the experimental results on ozone absorption given by Walshaw (1957) than are those which retain the P/P_0 factor.

For all gases at high altitudes, where the absorber concentration is low, most of the lines

which form an absorption band have not yet reached total absorption in the center of the lines and absorption is linearly proportional to the amount of gas and independent of pressure. Therefore, the pressure and temperature correction applied at these heights introduces a small error. The effect of Doppler broadening on lines was also neglected, but should be considered for heights greater than about 50 km (Plass and Fivel, 1953: see also Goody, 1954). We believe, however, that these effects are relatively unimportant compared to the uncertainties in the temperature profiles and distributions of water vapor and ozone in the higher atmosphere.

The computations were performed for a concentrically stratified atmosphere (see Fig. 1), the distributions of absorbing gases, clouds, temperatures, and pressures then being functions of height only.

Since $du=qds$, the reduced optical path, u^*, corrected for pressure and temperature is given by

$$u^* = +\int_{s=-\infty}^{s} \frac{P(s)}{P_0}\left(\frac{T_0}{T(s)}\right)^{1/2} q(s)\, ds$$

$$= -\int_{r=+\infty}^{r} \frac{P(r)}{P_0}\left(\frac{T_0}{T(r)}\right)^{1/2} q(r)\frac{r}{(r^2-r_0^2)^{1/2}}\, dr. \quad (3)$$

The integration along the path s has to be performed from the "top" of the atmosphere to a level r but not farther than to the point where the particular ray comes closest to the surface (e.g., for $r_0 < R$, to the surface, and for $r_0 > R$, to Point A in Fig. 1).

The refraction of rays in the density gradient has been neglected. Although the refraction of individual rays in the density gradient may not be entirely negligible, in the problem of horizon sensing this effect would to the first approximation cancel because of symmetry about the local vertical. For this reason, we neglected refraction in this investigation. For $r_0=0$ we obtain the value of u^* required for the calculation of the radiance leaving the atmosphere in the direction of the radius vector. Equation (3) can be used also to investigate the limb darkening effect as a function of the zenith angle θ, since $r_0=r\sin\theta$. The radiation chart used easily permits the second symmetrical half of the atmosphere (to the right of A in Fig. 1) to be taken into account. The radiation chart also deals in a simple manner with clouds and the surface of the Earth, and therefore these parameters do not have to be considered here in the computation u^*.

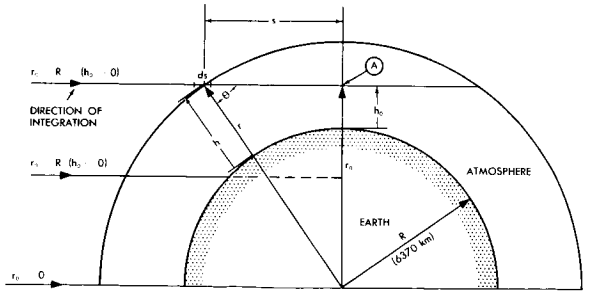


FIGURE 1.—Geometry for the calculations.

The pressure $P(h)$ chosen for a particular model atmosphere was computed separately by numerically integrating the hydrostatic equation

$$P(h) = P_0 \exp -\int_0^h \frac{g(h)}{R_d T(h)}\, dh. \quad (4)$$

The gas constant for dry air R_d, has been taken as 2.8704×10^6 cm^2 sec^{-2} (°K)$^{-1}$. (The effect of atmospheric water vapor on the mean molecular weight of air has been neglected; altitude corrections for $g(h)$, although small, have been made.) The function $q(r)$ in (3) gives the distribution of the absorbing gas. For CO_2 we assumed a homogeneous distribution of 0.00031 fractional parts by volume. Consequently du along the radius vector dr becomes

$$du \text{ (cm NTP)} = q_{CO_2}(r)\, dr = 31(\rho_{air}/\rho_0)\, dr \text{ (km)}. \quad (5)$$

But since $\rho_{air} = P/R_dT$

$$du \text{ (cm NTP)} = 31(P/P_0)(T_0/T)\, dr \text{ (km)}$$

and

$$u^*_{CO_2} \text{ (cm NTP)} = -31\int_{\infty}^{r}\left(\frac{P(r)}{P_0}\right)^2 \times\left(\frac{T_0}{T(r)}\right)^{3/2}\frac{r}{(r^2-r_0^2)^{1/2}}\, dr \text{ (km)}. \quad (6)$$

Similar considerations for water vapor and ozone yield

$$u^*_{H_2O}\ (\text{cm prec } H_2O) = -0.1292\int_{\infty}^{r}\left(\frac{P(r)}{P_0}\right)^2 \times\left(\frac{T_0}{T(r)}\right)^{3/2} q_{H_2O}(r)\frac{r}{(r^2-r_2^2)^{1/2}}\,dr\ (\text{km}), \quad (7)$$

$$u^*_{O_3}\ (\text{cm NTP}) = -\int_{\infty}^{r}\frac{P(r)}{P_0} \times\left(\frac{T_0}{T(r)}\right)^{1/2} q_{O_3}(r)\frac{r}{(r^2-r_0^2)^{1/2}}\,dr\ (\text{km}). \quad (8)$$

For O_3, where $(P/P_0)\geq 0.132$, a constant pressure ratio $(P_1/P_0)=0.132$ was used. The water vapor mixing ratio q_{H_2O} is expressed in grams per kg and q_{O_3} is given in cm NTP per km. (The distinction between specific humidity and mixing ratio of water vapor was neglected in these calculations.) However, before the numerical calculations of u^* could be carried out, a number of model atmospheres had to be selected, and this is discussed next.

THE SELECTION OF MODEL ATMOSPHERES

The model atmospheres were designed to illustrate the range of radiation characteristics which might reasonably be expected in the various spectral ranges considered. Accordingly, emphasis has been placed for the most part on the extremes likely to be found in various atmospheric parameters. However, in all cases an effort has been made to keep the models plausible and internally consistent.

Five different temperature profiles were used (Fig. 2). These are intended to cover the latitude variations and, in the case of the arctic atmosphere, the seasonal variations. The profiles are:

1. *Tropical atmosphere.* The tropical profile given in the *Handbook of Geophysics* (1960) was used up to approximately 22 km. Above this level the low-latitude standard profile proposed by Nordberg and Stroud (1961) was used (determined by sounding rocket measurements on the island of Guam).

2. *Desert atmosphere.* A linearized average of two desert soundings was used up to 30 km, with the Norberg–Stroud (1961) low latitude standard being incorporated above 30 km. This profile is typical of low-latitude desert regions.

3. *ARDC standard atmosphere* (1959). This profile (*Handbook of Geophysics*, 1960) is typical of the mid-latitude regions.

4. *Arctic summer atmosphere.* The arctic summer profile given by the *Handbook of Geophysics* (1960) was used up to 30 km. Above 30 km the high-latitude summer profile given by Nordberg and Stroud (1961) was adopted. This combination was chosen to give the warmest arctic stratosphere.

5. *Arctic winter atmosphere.* The arctic winter profile as given in the *Handbook of Geophysics* (1960) was used up to 24 km. Above 24 km, the high latitude winter profile given by Nordberg and Stroud (1961) was chosen. This combination was adopted to show the influence of a cold arctic stratosphere.

The distributions of the optically active gases were chosen, as were the temperature profiles, to illustrate the range of conditions that might be expected. The proportion of carbon dioxide in the atmosphere was taken constant with respect

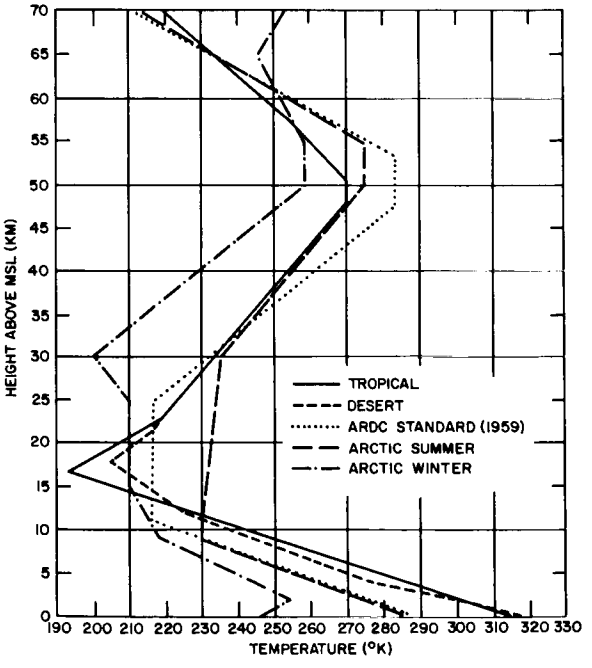


FIGURE 2.—Model atmosphere temperature profiles.

to altitude and latitude, and equal to 0.031 percent by volume. The water vapor mixing ratios adopted are shown in Fig. 3 and described below:

1. *Tropical model.* Up to 8 km, mixing ratios based on a linearization of a Trinidad, West Indies, sounding (1200 UT, 20 July 1961) were used. This sounding was chosen as representative of a very moist tropical troposphere. From 8 to 20 km a constant relative humidity was assumed. From 20 to 31 km a constant frost point was used, and above 31 km a constant mixing ratio of 0.11 g per kg. This combination results in a total of about 5.07 cm of precipitable water. Temperature profile (1) was incorporated in this model.

2. *Desert model.* A linearized average of two Aoulef, Algeria, soundings (1200 UT, 16 August 1960, and 1200 UT, 25 August 1960) was used up to 17.5 km. From 17.5 km to 21 km a constant frost point was assumed, and above 31 km the mixing ratio was held constant at 0.06 g per kg. This results in a total of 0.9 cm precipitable water.

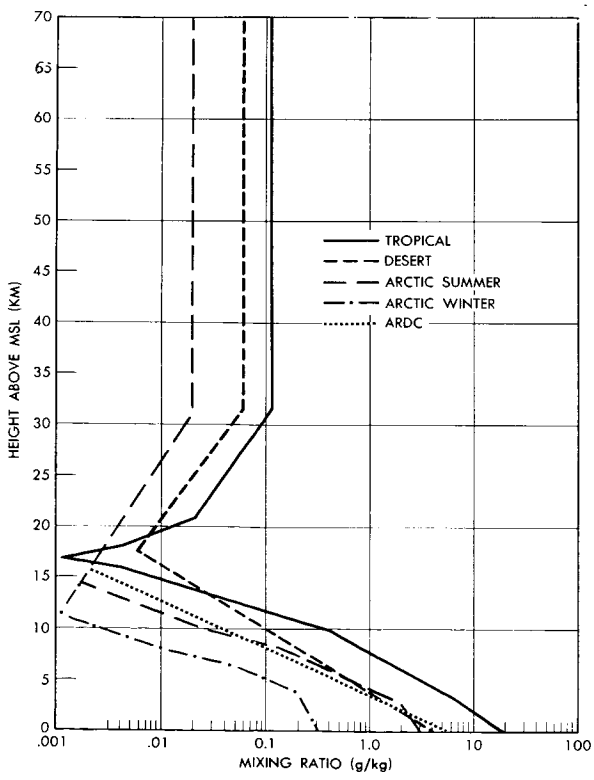


FIGURE 3.—Model atmosphere water vapor mixing ratios. Above about 15 km, the ARDC and arctic winter mixing ratios are coincident with those of the arctic summer.

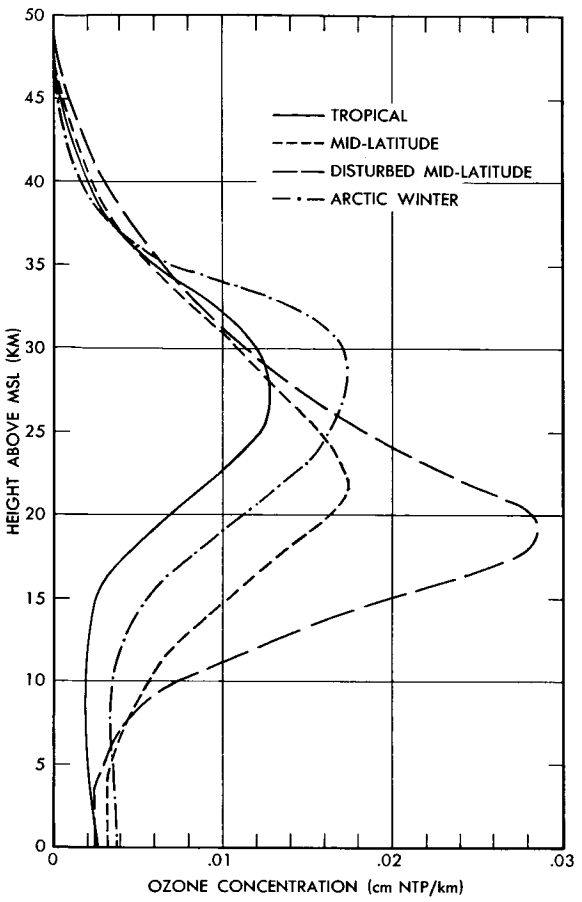


FIGURE 4.—Model atmosphere ozone distribution.

3. *Arctic summer model.* The water vapor mixing ratios were taken from the altitude and latitude distributions of water vapor given by Manabe and Möller. (1961).

4. *Arctic winter model.* The mixing ratios for this atmosphere were taken from Manabe and Möller (1961).

5. *ARDC standard atmosphere* (1959). A representative mixing ratio profile was adopted, resulting in a total of 1.16 cm of precipitable water.

The four ozone models chosen are shown in Fig. 4 and described below:

1. *Tropical model.* Temperature profile (1) was employed along with the ozone distribution for 5N latitude used by Manabe and Möller (1961).

2. *Mid-latitude model (Arosa normal).* An ozone distribution based on a three-year mean (1956–1958) of measurements made at Arosa, Switzerland (47N), (Dütsch, 1959), was chosen along with temperature profile (3).

3. *Disturbed mid-latitude model (Arosa high).* This model was constructed as an illustration of the rapid changes that might be encountered at at any given point as the result of the passage of frontal systems. The ozone distribution used is that given by Dütsch (1959) based on measurements taken after the passage of a pronounced cold front. As in the Arosa normal case, temperature profile (3) was used.

4. *Arctic winter model.* Temperature profile (5) was used in this model along with the Manabe and Möller (1961) distribution of ozone for January at 85N.

NUMERICAL COMPUTATION OF u^*

A large number of values of u^* had to be calculated for each case to allow curves of u^* versus τ to be plotted on the radiation diagrams. For these calculations (6), (7) and (8) were solved for carbon dioxide, water vapor and ozone, respectively, on an IBM 7090 Computer. The integration over r was carried out by a straight summation using intervals of 0.05 km and initiated at $h_{max} = 70$ km.

CONSTRUCTION OF RADIATION CHARTS

A radiation diagram was constructed for each of the five spectral regions chosen. The various types (Möller, Elsasser, Yamamoto, etc.) are similar in their basic concept and can be transformed into each other by conformal mapping. We adopted Möller's (1951) version of the diagram and Elsasser's (1960) generalized absorption coefficients. For our computations, we divided each region into a number of spectral intervals for which we assumed a constant absorption coefficient. The values adopted are summarized in table 2.

The radiation diagrams for the carbon dioxide band, the window, the ozone band and the 6.3-micron water vapor band are comparatively simple to construct since the wavelength regions are so small that the distribution of blackbody energy,

TABLE 2.—*Absorption coefficients adopted for the calculations*

Major absorber	Wavelength (microns) λ	Band absorption $\tau = \tau(Lu^*)$ $\log L$	Absorption by continuum $\tau = \exp(-ku^*)$ $\log k$
H_2O	6.33–6.5	+2	—
	6.5–6.67	0.4+2	—
	6.67–6.85	+2	—
O_3	8.9–9.35	0.8−2	—
	9.35–9.9	0.2	—
	9.9–10.1	0.6−1	—
Window (H_2O continuum)	10.75–11.75	0.7−5	0.85−2
CO_2	14–14.5	0.6−1	—
	14.5–15.5	0.3	—
	15.5–16	0.6−1	—
H_2O	21–25	+1	—
	25–30	0.6+1	—
	30–40	0.2+2	—
	40–125	0.6+2	—

relative to the spectral intervals into which each region has been divided (table 2), remains approximately constant over the temperature range of interest. Therefore, lines of constant temperature are horizontal.

This simplification, however, does not hold for the broad wavelength region being considered within the rotational band of water vapor. For the computation of this radiation diagram, the integral

$$\int_{\lambda} B_{\lambda}(T) \frac{\partial \tau}{\partial u^*} \, d\lambda \approx \sum_i B_{\Delta\lambda_i}(T) \frac{\Delta \tau_i}{\Delta u^*}, \qquad (9)$$

which is necessary in the construction of the ordinate, was approximated by a summation over four wavelength intervals shown in table 2.

Here $B_{\Delta\lambda_i}$ indicates that the actual fraction of the Planck function bounded by the somewhat broad limits of each of the four wavelength intervals is used, rather than the product of the spectra blackbody radiant emittance at the central wavelength and the wavelength interval which was used in the numerical calculations for all of the other regions.

In the window region, the absorption is mainly determined by the absorption coefficient k. The numerical value used in our computation is a factor of 10 higher than the one given by Elsasser (1960) which apparently should be used with $10\,u^*$. In all computations, the generalized absorption coefficient was considered to be independent of temperature. The scales in all diagrams, were calculated with the aid of the radiation slide rule from the Admiralty Research Laboratory, Teddington, England.

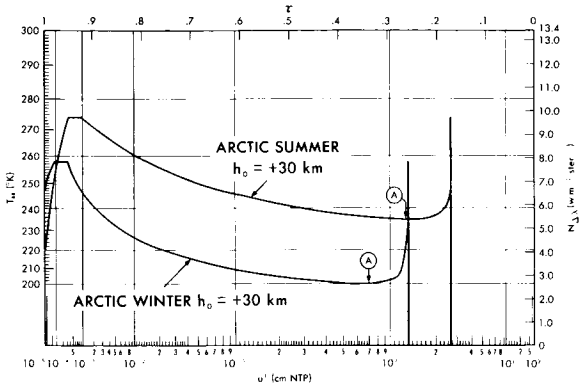


FIGURE 5.—Radiation chart for the 15-micron CO_2 band. Radiance values correspond to the area under each curve. The right hand portion of each curve is an image of the plot to the left of point A. The lower arctic winter radiance is caused by lower temperatures and lower air densities in the upper atmosphere.

EVALUATION OF THE RADIATION CHARTS

The atmospheric models were plotted on the radiation charts and the integrations performed graphically using a planimeter. In the plotting procedure, the part of the atmosphere beyond the closest approach to the Earth (to the right of point A in Fig. 1) was taken into account by imagining the temperature profile in the radiation diagram at the u^* value corresponding to $r = r_0$ (Fig. 5). In cases where the atmosphere was transparent enough to make the surfaces contribute to the outgoing radiation, the surface was assumed to radiate as a blackbody at the temperature indicated by the lowest point of the particular temperature profile. This effect occurs, for example, in the window case for the arctic winter atmosphere (Fig. 6).

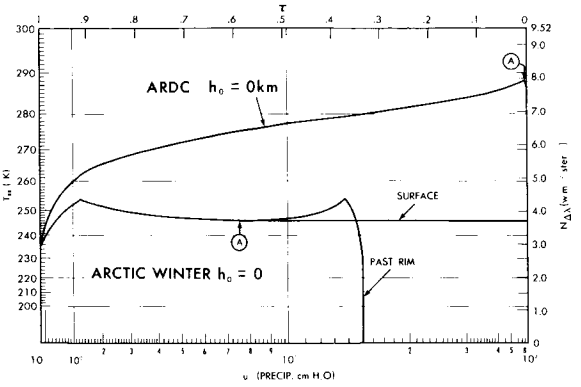


FIGURE 6.—Radiation chart for the atmospheric window. When the atmosphere is not opaque, two solutions exist for $h_0 = 0$, viz., when the ray is terminated at the surface and when the ray extends past the rim of the earth.

In addition to the clear-sky states of the model atmospheres, high cloud overcasts were introduced into all except the desert atmosphere. The clouds, too, were assumed to radiate as blackbodies at the ambient temperatures associated with the heights of cloud tops. To illustrate extreme cases, the cloud-top heights have been considered to be at the tropopause (and, therefore, at the lowest temperature) for the tropical atmosphere (16.8 km), as shown in Fig. 7, and for the arctic summer atmosphere (9.0 km). For the standard atmosphere, the cloud-top height has been placed in the minimum temperature isothermal region at 12.0 km, a representative level considering the variation in the height of

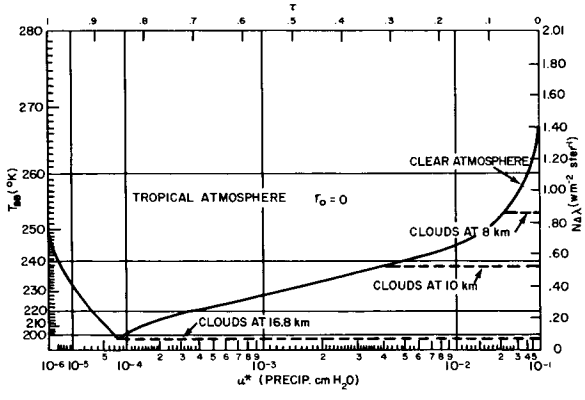


FIGURE 7.—Radiation chart for the 6.3-micron H_2O band. The effect of clouds acting as blackbody radiators at various heights in the tropical atmosphere is illustrated.

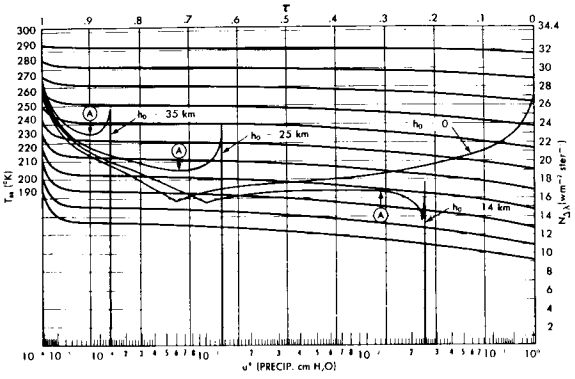


FIGURE 8.—Radiation chart for the rotational H_2O band. Curves are plotted on the desert atmosphere for various values of h_0.

the tropopause over middle latitudes. Considering the often ill-defined tropopause at high latitudes in the winter, the cloud top height has been placed at 8.5 km for the arctic winter atmosphere. This level is near the mean winter tropopause height at high latitudes according to Kochanski (1955).

The radiation chart of the rotational water vapor band (Fig. 8) shows samples of the desert atmosphere. The decrease of N (area under each curve) with rising values of h_0 is caused by a rapid reduction in the absorbing gas. This rapid decrease is partially compensated by a simultaneous increase in radiance per unit $\Delta\tau$ as a result of the positive, vertical temperature gradient in the stratosphere and the shape of the isothermal lines in the radiation chart. The point A, for example, rises 20 deg from the curve corresponding to $h_0=25$ to the one corresponding to 35 km. The high transparency of the zone above 25 km shifts the weight of the emitted radiation towards the range which possesses the highest absorption coefficient, i.e., towards the long wavelength end of the 21- to 125-micron interval (see table 2). There the Planck function is less sensitive to small temperature changes, which is the physical reason for the rise of the isothermal lines at the left side of the radiation chart.

Special attention must be given to the second important absorber within each wavelength region. Only in the range between 8.9 and 10.1 microns was it necessary to consider two gaseous absorbers quantitatively, namely, ozone and the water vapor continuum. This was done, as shown in Fig. 9, by plotting the calculated transmissivity curve $\tau=\tau_{O_3}\cdot\tau_{H_2O}$ and integrating over it. The absorption coefficients for the window (10.75 to 11.75 microns) were used to compute τ_{H_2O}.

The influence of water vapor on the 14–16 micron CO_2 emission band can be neglected because of the opaqueness of CO_2 ($\tau_{CO_2}\leq 0.02$) at a level where water vapor is still transparent ($\tau_{H_2O}\geq 0.98$). A similar consideration for the effect of the 14-micron ozone band within the 14–16 micron range is not as favorable. In the case of the standard atmosphere, for an h_0 of zero, the ozone diagram for 14–16 microns shows a transmissivity from the surface of only about 0.8. However, the emission originates mostly from the 10–25 km low temperature range which also dominates the carbon dioxide diagram. The net effect of the residual ozone emission is then the same as the CO_2 emission and it was therefore neglected.

The window from 10.75 to 11.75 is so transparent that small amounts of solid particles or droplets in the air affect transmissivity. For example cirrus clouds, often invisible from the surface, or dust trapped near the tropopause seem to have more influence on the outgoing radiation in the window than commonly thought.[2,3]

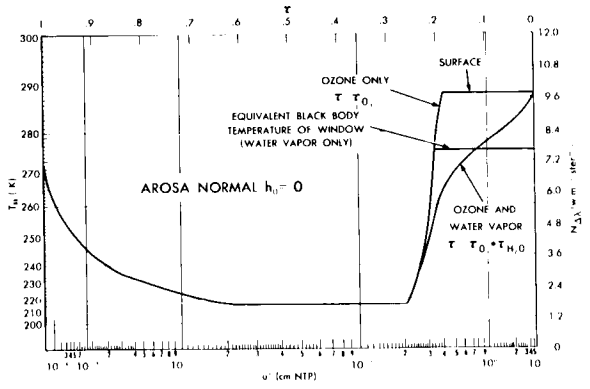


FIGURE 9.—Radiation chart for the 9.6-micron O_3 band. Three curves are illustrated: (1) where the transmission is due to O_3 only; (2) where the first curve is terminated at the equivalent blackbody temperature of the window (considering the entire atmosphere); and (3) where the transmission is taken as the product of the transmissions due to both O_3 and H_2O. It is seen that radiance values calculated from the latter two curves are approximately equal.

We have neglected this issue since very little is known about the altitude distribution of aerosols and dust. Its effect would be a reduction in the radiance emitted from the Earth reaching satellite altitudes and a possible increase of radiance in the clear atmosphere for h_0 values between 5 and 20 km.

RESULTS OF THE INTEGRATION

The results of the integration are summarized, for each wavelength interval, in Figs. 10 through 14. In the diagrams, the radiance N ($w\ m^{-2}$ ster^{-1}) within the spectral band and the corresponding blackbody temperature (°K) are plotted versus altitude h_0 above or below the geometrical horizon ($h_0=0$). Included at the left side of each figure is the radiance value for $r_0=0$, which is the radiance emerging from the Earth in the direction of the radius vector.

A comparison of the radiance for $r_0=0$ (center of disc) and $h_0=0$ (horizon) shows a small but definite limb brightening effect for all atmospheres in the 14–16 micron band (Fig. 10). Almost all calculated model atmospheres show a maximum in the vicinity of +20 km, and an almost identical behavior above 30 km. The arctic winter atmosphere has the same general appearance, but deviates somewhat in the absolute levels. The reasons for this are the low temperatures of the upper atmosphere and the associated low densities, which are consistent with the low temperatures

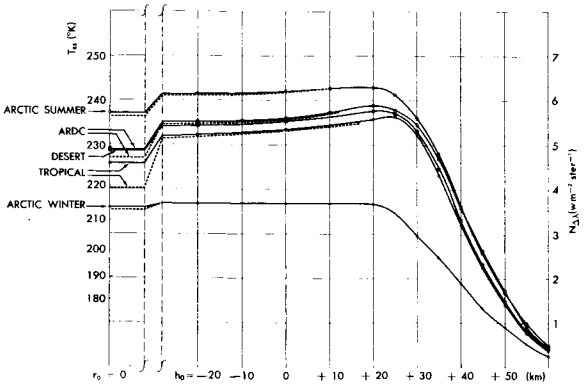


FIGURE 10.—Curves of radiance in the 15-micron CO_2 band vs. h_0 (r_0) for the five model atmospheres. Solid lines represent clear sky and dashed lines represent high cloudiness conditions.

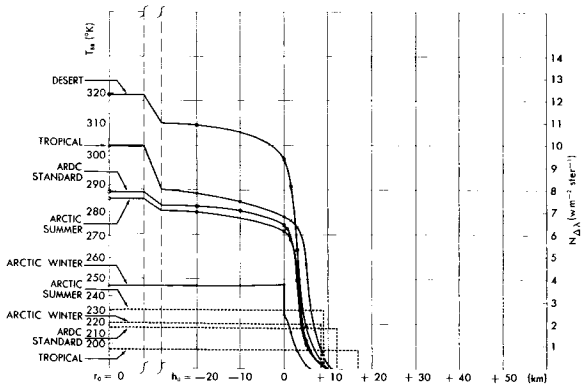


FIGURE 11.—Curves of radiance in the atmospheric window vs. h_0 (r_0) for the five model atmospheres. Solid lines represent clear sky and dashed lines represent high cloudiness conditions.

in the troposphere. Both effects can be observed in Fig. 5. Clouds have a noticeable influence on the radiance in the tropics only, and there just near the center of the disk. In the strong CO_2 band, the Earth appears very uniform with the near the center of the disk. In the strong CO_2 band, the Earth appears very uniform with the maximum radiance near the rim and with a generally warmer summer and cooler winter pole.

The appearance of the Earth in the atmospheric window (see Fig. 11) is quite different from that seen in the CO_2 region; the slope of the horizon is generally much steeper. Only in the arctic winter case does the geometrical horizon become visible. In practice, this will occur rather rarely, since, for the most part, clouds will obscure the view, making the disc of the Earth appear rather irregular. Enormous changes in the radiance levels from one point of the horizon to an adjacent one must be expected, e.g., deserts emitting about 10 $w\ m^{-2}$ ster^{-1} may be next to high clouds emitting only 0.9 w m^{-2} ster^{-1}. The meteorological situation has a great influence on this wavelength region and the threshold of a horizon detector system must be set low enough that high cloudiness will not be mistaken for outer space.

The 6.3-micron water vapor band shows less contrast than the window at the center of the disc ($r_0=0$) (Möller 1961) and even less near the horizon (Fig. 12). The low temperatures at the tropical tropopause and an increase in the water vapor mixing ratio in the stratosphere give rise

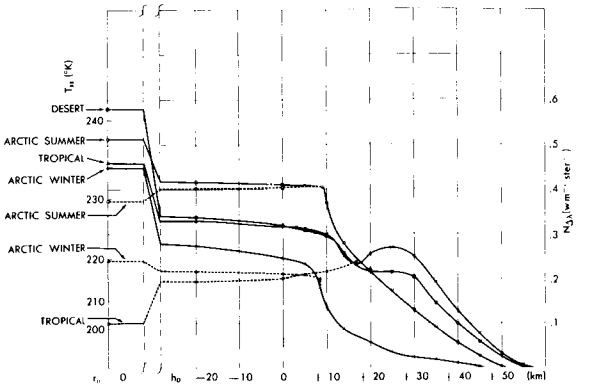


FIGURE 12.—Curves of radiance in the 6.3-micron H_2O band vs. h_0 (r_0) for four model atmospheres. Solid lines represent clear sky and dashed lines represent high cloudiness conditions.

to a warm zone at 25 km above a partial minimum near 15 km. The same effect creates a small plateau in the desert atmosphere between 20 and 30 km. Tropical clouds, however, are still visible near the horizon.

The general shape of the horizon in the rotational water vapor band, shown in Fig. 13, appears much more favorable, although the absorption coefficients are of the same order of magnitude as in the 6.3-micron band. The reason for this lies in the energy distribution of the Planck function; the radiance in the 6.3-micron band is proportional almost to the 10th power of temperature within the temperature limits of 200° and 250°K, whereas the rotational water vapor band in this same region is proportional to a power of only two or three. The contrast between radiance values across the disc of the Earth is therefore much less than in the 6.3-micron band. In the water vapor band, all atmospheres show a steep decline at about a 10-km altitude. Tropical clouds again are visible and minor humps occur similar to those in the other water vapor band at 6.3 microns.

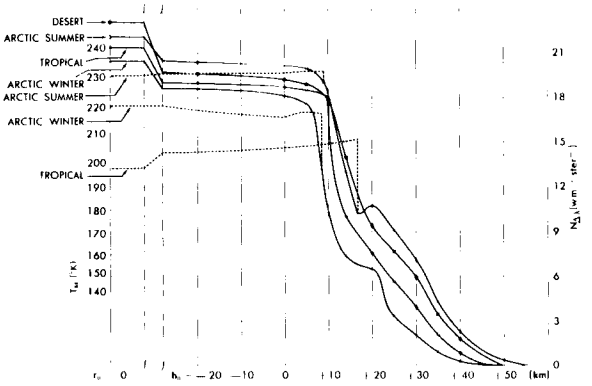


FIGURE 13.—Curves of radiance in the rotational H_2O band vs. h_0 (r_0) for four model atmospheres. Solid lines represent clear sky and dashed lines represent high cloudiness conditions.

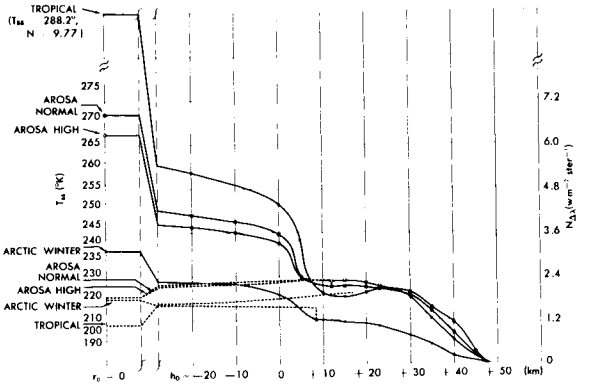


FIGURE 14.—Curves of radiance in the 9.6-micron O_3 band vs. h_0 (r_0) for the tropical, "Arosa normal," "Arosa high," and arctic winter atmospheres. Solid lines represent clear sky and dashed lines represent high cloudiness conditions.

In the ozone absorption region, shown in Fig. 14, the clear atmosphere exhibits a very pronounced limb-darkening effect. Similar to the window region, strong contrasts between clear and cloudy areas exist. The stratification of ozone yields a flat minimum or a small plateau between 10 and 30 km, which is analogous to the situation for tropical water vapor. One can even speak of a double horizon, one near 40 km and the second one near 5 km, the latter one appearing only under clear sky conditions. A comparison of the "Arosa" ozone distributions shows a small but interesting phenomenon. An increase of about 50 percent in the ozone content in the upper atmosphere decreases the radiance of the disc (only by a small amount, perhaps 8 percent), but causes a rise in radiance for h_0 values above 5 km. The arctic winter atmosphere is also exceptionally low in radiance. Again in the ozone region clouds contrast strongly against clear skies.

COMPARISON OF CALCULATED VALUES TO TIROS RADIATION DATA

The TIROS meteorological satellite radiation experiment provides observational data for comparison with some of the above calculations. One channel of the radiometer was sensitive to radiation in the 8–13 micron region, and it is therefore possible to make a rough comparison with this observed region by combining the calculations for the 10.75–11.75 micron and 8.9–10.1 micron bands. While the five degree field-of-view of the TIROS radiometer does not permit sufficiently high resolution of the horizon, it is possible to compare data taken with the sensor axis nadir angle near zero with the calculations made for $r_0 = 0$.

It is assumed that the radiance per wavelength increment seen in the 8–13 micron region would be the same as that seen in the 10.75–11.75 micron region, with the exception of the 8.9–10.1 micron ozone band, where the calculated radiance for that region was used. In this way the total blackbody temperature was derived. The temperature thus calculated for the standard atmosphere and clear skies was 280°K compared, for example, to an average of 289°K for clear areas over the eastern United States (Nordberg et al., 1962). The calculated temperature for the tropical atmosphere and clouds near the tropopause of 201°K compares favorably with the observed temperature over Hurricane Anna of about 204°K (Bandeen et al., 1962).

The above comparisons show reasonable agreement, considering the strong dependence of the radiance on the temperature profile and gas distributions. A more thorough comparison of observation and theory would necessitate soundings of temperature and gas distributions at the times the satellite data were acquired. Calculations could then be made, using the sounding data, and could be compared with the radiation observations.

CONCLUSIONS

For the purpose of horizon sensing, the choice of a spectral region is governed by a number of considerations. One important design parameter is the lowest value of radiance, $N_{\min}$, within the disk of the Earth. A horizon sensor must identify this level as Earth in contrast to space. Since, in certain design concepts, the detector receives radiation only from a small region near the horizon and never from the central part of the disk, the minimum radiance for $h_0 = 0$ is important. In table 3, a value N_0 slightly below the minimum radiance is listed which could serve as a conservative estimate of this threshold value. The difference, Δh, between the highest and lowest horizon, quoted for the threshold level N_0, is important only for sensors which are capable of resolving the true shape of the horizon.

The curves in Figs. 10 through 14 may also be interpreted as time functions obtained from an ideal detector of zero time constant, without noise, and with an infinitely narrow field of view scanning across the horizon. A more realistic time-constant-to-scan speed ratio and a finite field of view can be considered to be bandwidth restrictions (electrical filters) applied to the Fourier spectrum of the horizon functions. The influence of detector noise on horizon determinations was discussed recently by Goetze (1962).

For many instrument designers the ratio between the maximum and the minimum value of N becomes very significant; these ratios are listed for the center of the disk ($r_0 = 0$) and for the horizon ($h_0 = 0$).

The great advantages of the long-wavelength, high-absorption regions between 14 and 16 and beyond 21 microns are apparent. The radiance values are higher, uncertainties in the horizon (Δh) caused by meteorological and seasonal effects are lower, and the contrast across the disk of the Earth is low in comparison with the values in the shorter wavelength regions including the window. Unfortunately, instrumental techniques for wavelength longer than 30 or 40 microns are not as highly developed as those at shorter wavelengths. However, the 15-micron CO_2 band is well within the convenient operating range of thermal detectors, and filters are commercially available.

The comparison of spectral regions in table 3 includes seasonal effects as well as local meteorological phenomena. It can be expected that the two effects enter accuracy considerations in a different manner. For example, the low arctic winter radiance in the 15-micron CO_2 band is

TABLE 3.—*Comparison of radiance of spectral regions*

Wavelength (microns)	N_{min} ($r_0 = 0$) (w m^{-2} ster^{-1})	N_{min} ($h_0 = 0$) (w m^{-2} ster^{-1})	N_0 (w m^{-2} ster^{-1})	Δh (km)	N_{max}/N_{min} ($r_0 = 0$)	N_{max}/N_{min} ($h_0 = 0$)
6.33– 6.85	0.102	0.201	0.08	29.2	5.69	2.05
8.9 –10.1	1.09	1.57	0.9	14.7	8.95	2.79
10.75–11.75	0.86	0.86	0.7	13.4	14.3	10.9
14 –16	3.54	3.69	3.0	13.3	1.61	1.68
21 –125	13.3	14.7	12	7.8	1.74	1.37

semistationary and could, at least theoretically, be compensated for in the same manner as one could compensate for the equatorial bulge of the Earth. Meteorological effects on the other hand are more or less random and cannot be compensated for, other than by integration over time and space. From these considerations, the 15-micron band becomes even more attractive.

Considering both the theoretical and instrumentational aspects of the problem, it is concluded that a scanner sensitive in the interval of approximately 14 to 16 microns is most promising for minimal erros in the determination of the local vertical from a satellite by means of a horizon sensing system.

It is planned to include a 14- to 16-micron scanner on one of the TIROS satellites in the near future in order to compare theoretical with observed results in this spectral interval.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Fritz Möller of the University of Munich and Dr. D. Q. Wark of the U. S. Weather Bureau for many helpful discussions concerning the contents of this paper.

Also, the authors should like to express their appreciation to Mr. Robert Hite and Mr. Oliver Clark for programming the numerical computations on the computer and to Miss Louise Coury, Mr. William Brastow, Jr., and Mr. Terry Mahuron for their painstaking and careful plotting and graphical evaluation of the radiation charts.

TABLE OF SYMBOLS

$B_\lambda(T)$ Spectral radiant emittance (Planck function) for a blackbody of temperature T (watts m^{-2} micron^{-1})

g acceleration due to gravity

h height above the surface of the earth

h_0 height of line s (along radius vector normal to s)

L generalized absorption coefficient

N infrared radiance (watts m^{-2} ster^{-1})

NTP conditions of normal temperature (T_0) and pressure (P_0)

P atmospheric pressure

P_0 normal pressure (760 mm Hg, 1013.25 mb)

P_1 pressure of 10 cm Hg, above which the pressure correction factor P/P_0 for calculating the u^* for ozone remains constant at $P_1/P_0 = 0.132$

R radius of assumed spherical earth (6370 km)

R_d gas constant for dry air (2.8704×10^6 cm^2 sec^{-2} deg^{-1})

r geocentric distance

r_0 geocentric distance to line s (length of radius vector normal to s)

s straight line path along which the integration of u^* is carried out

T kinetic temperature

T_{BB} equivalent blackbody temperature corresponding to integrated radiance within a given spectral interval

T_0 normal temperature (0°C, 273.16°K)

u actual optical thickness of an absorbing (and radiating) gas

u^* reduced optical thickness, corrected for pressure and temperature

θ zenith angle of s

λ wavelength (microns),

q distribution function of an optically active gas

ρ_{air} density of air

ρ_0 density of air at NTP

τ transmissivity for straight-line beam radiation

ϕ_λ spectral response of a sensor

REFERENCES

1. Goetze, D., 1962: Private communication, May.
2. Suomi, V., 1962: Private communication, Feb.
3. Möller, F., and Zdunkowski, 1962: Technical report. Computational methods of long wave atmospheric ratiation. Contract AF 61(052)–493, 49 pp.

GENERAL REFERENCES

BANDEEN, W. R., B. J. CONRATH, W. NORDBERG AND H. P. THOMPSON: A radiation view of Hurricane Anna from the TIROS III meteorological satellite. To be published in *Proceedings of the International Symposium on Rocket and Satellite Meteorology, Washington, D. C., April 23–25.* Amsterdam, North-Holland Publishing Company, 1962.

CONRATH, B. J.: Earth scan analog signal relationships in the TIROS radiation experiment and their application to the problem of horizon sensing. *NASA Technical Note D-1341*, Goddard Space Flight Center, Greenbelt, Md., 16 pp. (Available from Office of Scientific and Technical Information, Code AFSC, NASA HQ, Washington 25, D. C.), 1962.

DÜTSCH, H. U.: Vertical ozone distribution from Umkehr observations. *Archiv für Meteorologie, Geophysik und Bioklimatologie*, **11**, 240–251, 1959.

ELSASSER, W. M.: Atmospheric radiation tables. *Meteorological monographs*, Vol. 4, No. 23, Boston, Amer. Meteor. Soc., 43 pp., 1960.

GOETZE, D.: Accuracy and range of infrared horizon sensors as limited by detector noise. *Amer. Rocket Soc. J.*, **32**, 1039–1044, 1962.

GOODY, R. M.: *The physics of the stratosphere.* Cambridge, University Press, 187 pp., 1954.

Handbook of geophysics: Revised edition. Air Research and Development Command, Air Force Research Division, Geophysics Research Directorate, New York, The Macmillan Company, 656 pp., 1960.

KOCHANSKI, A.: Cross sections of the mean zonal flow and temperature along 80W. *J. Meteor.*, **12**, 95–106, 1955.

KONDRATIEV, K. Y., and K. E. YAKUSHEVSKAYA: The angular distribution of the outgoing thermal radiation in the different regions of the spectrum. To be published in *Proceedings of the International Symposium on Rocket and Satellite Meteorology, Washington, D. C., April 23–25.* Amsterdam, North Holland Publishing Company, 1962.

MANABE, S., and F. MÖLLER: On the radiative equilibrium and heat balance of the atmosphere: *Mon. Wea. Rev.*, **89**, 503–532, 1961.

MÖLLER, F.: Long-wave radiation. *Compendium of meteorology*, Boston, Amer. Meteor. Soc., 34–49, 1951.

———: Atmospheric water vapor measurements at 6–7 microns from a satellite. *Planet. Space Sci.*, **5**, 202–206, 1961.

NORDBERG, W., and W. G. STROUD: Seasonal, latitudinal, and diurnal variations in the upper atmosphere. *NASA Technical Note D-703*, Goddard Space Flight Center, Greenbelt, Md., 16 pp. (Available from Office of Scientific and Technical Information, Code AFSC, NASA HQ., Washington 25, D. C.), 1961.

NORDBERG, W., W. R. BANDEEN, B. J. CONRATH, I. PERSANO, and V. KUNDE: Preliminary results of radiation measurements from the TIROS III meteorological satellite. *J. atmos. Sci.*, **19**, 20–30, 1962.

PLASS, G. N., and D. I. FIVEL: Influence of Doppler effect and damping on line absorption coefficient and atmospheric radiation transfer. *Astrophys. J.*, **117**, 225–233, 1953.

WALSHAW, C. D.: Integrated absorption by 9.6 micron band of ozone. *Quart. J. R. meteor. Soc.*, **83**, 315–321, 1957.

WORMSER, E. M., and M. H. ARCK: Guidance and control. *Proceedings of Guidance, Control and Navigation Conference*, (Stanford, Calif., August) New York, Academic Press, 443–468, 1962.

RELATION BETWEEN THEORETICAL AND OBSERVATIONAL MODELS OF THE UPPER ATMOSPHERE

I. HARRIS

Goddard Space Flight Center

AND

W. PRIESTER*

Bonn University Observatory

In a recent paper [Harris and Priester, 1962] we published theoretical working models for the solar-cycle variation of the upper atmosphere that were based on a working hypothesis concerned with the dependence of the heat sources on the solar activity. In that paper we assumed that the fluxes of both heat sources (extreme ultraviolet flux and 'corpuscular' heat flux) vary in proportion to the monthly averages of the solar 10.7-cm flux, which is generally used as an indicator of solar activity. Consequently the model parameters S of the theoretical models were taken to be equal to the monthly averages of the 10.7-cm flux.

In addition to the comparison made earlier with data for high solar activity, it is now possible to compare those models with observational data obtained during times of medium and low solar activity. This leads to an improvement of the working hypothesis. The observational results used are the models by King-Hele [1963] for the years 1958–1959, 1960, 1961, and 1962. An appreciably good agreement is found for all levels of solar activity if we take into account how large the density changes are during the solar cycle. Toward the lowest level of solar activity, however, a systematic deviation appears that yields an empirical relation between the monthly averages of the solar 10.7-cm flux $\bar{F}$ and the model parameters S which is shown in figure 1. This relation now replaces the working hypothesis used previously, and any user of our theoretical models should take the appropriate model according to this relation. The data given by King-Hele [1963] are yearly averages not corrected for the semiannual and annual

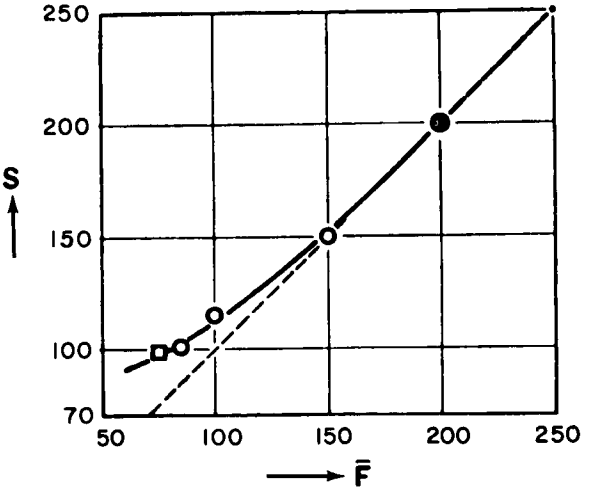


FIGURE 1.—Empirical relation between model numbers (Harris and Priester) S and the monthly averages of the solar 10.7-cm flux $\bar{F}$ as obtained from comparison with observational data: open circles, models by King-Hele for 1958–59, 1960, 1961, 1962; square, first data by Explorer 17, launched April 2, 1963; solid circle, Bonn model for $\bar{F}=200$. The dotted straight line represents the preliminary working hypothesis used in the paper by Harris and Priester [1962].

* Now at Goddard Institute for Space Studies, New York, as a NAS-NRC Senior Research Associate.

variation [Paetzold, 1963]. Our theoretical models, on the other hand, are generally applicable to average values for the months September through December, because the observational models by Martin et al. [1961], on which the theoretical models are based, apply to the average densities of the months given above. Therefore we can consider the curve in figure 1 as a lower limit. In particular, the open circles for $\bar{F}=150$ and 100

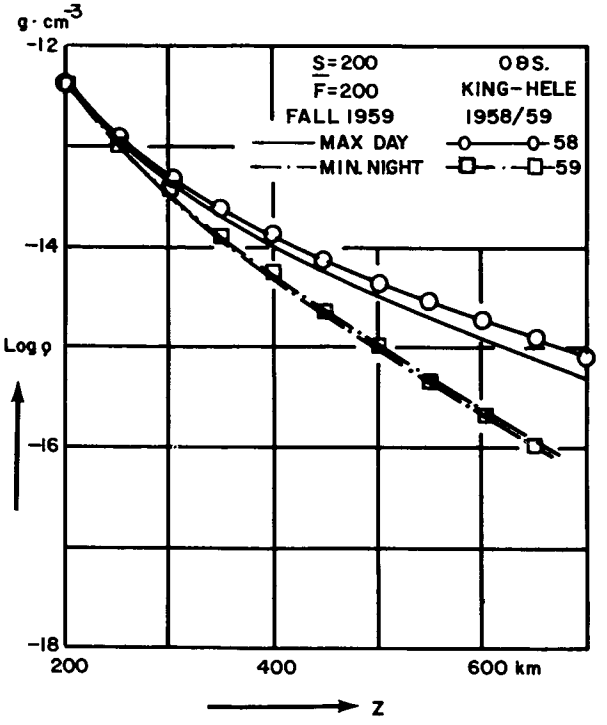


FIGURE 2.—Maximum daytime and minimum nighttime densities obtained by King-Hele from a large number of satellites are compared with the corresponding theoretical models by Harris and Priester. The observational curves are dashed at those altitudes where only a few data were available.

might be placed slightly too low. A more refined analysis is presently being carried out by Roemer [1963a,b].

Figures 2 to 5 show a comparison between the observational and theoretical models using the new relationship. The observational daytime maximum curve in figure 2 is for the year 1958 when the average of the solar 10.7-cm flux was 230. This accounts for the deviation from the theoretical model for an average flux of 200.

In figure 3 the merging of the observational

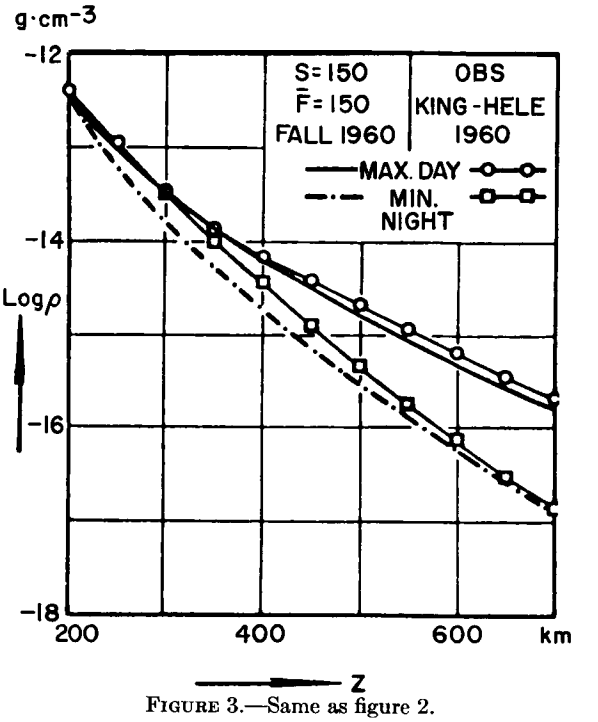


FIGURE 3.—Same as figure 2.

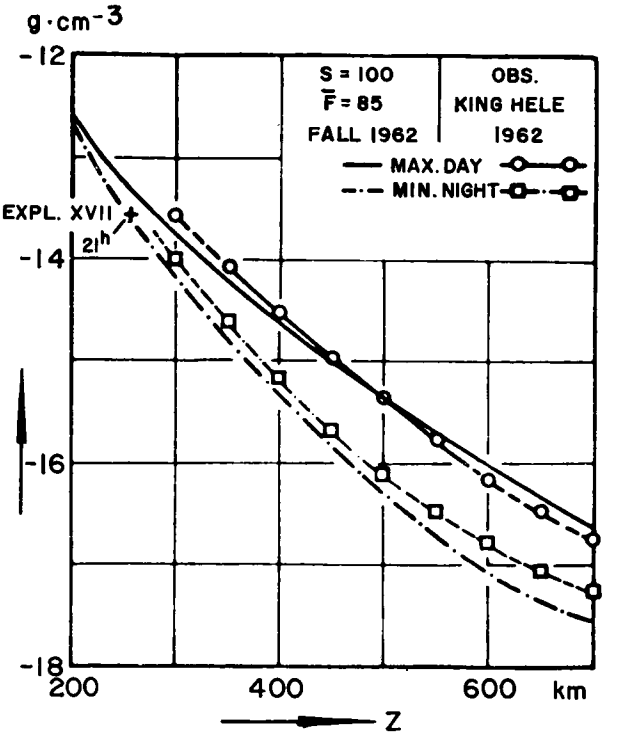


FIGURE 4.—Same as figure 2. The density measured on April 3, 1963, at 2100 local time by Explorer 17 is also given.

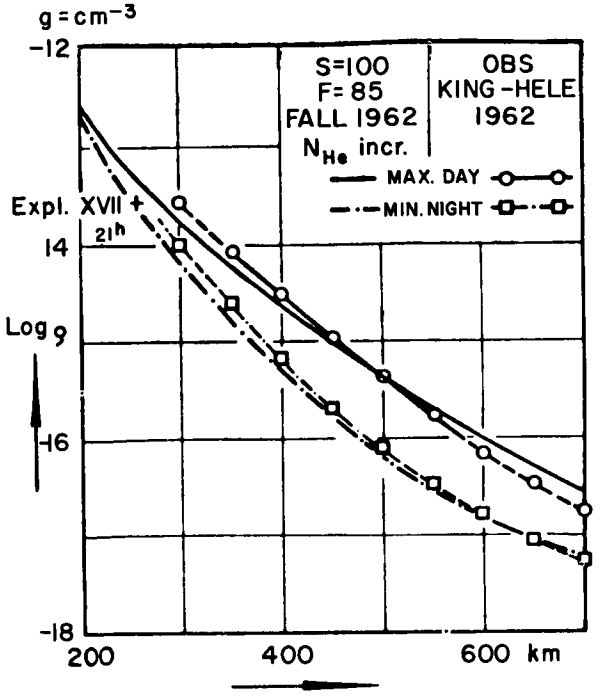


FIGURE 5.—Comparison between King-Hele's observed densities for 1962 and a new theoretical model containing an amount of helium increased by a factor 2.5 over that of the previous models. As in figure 4, the density obtained by Explorer 17 is also given.

curves for daytime and nighttime at 300 km is believed to be influenced by the method of extrapolation to lower altitudes. Theoretical considerations [Harris and Priester, 1962] make a crossover between day and night curve very unlikely at an altitude as high as 300 km. Furthermore, Jacchia and Slowey [1962] found an appreciably larger diurnal amplitude at 350 km from Explorer 1 for the year 1960.

In figure 4 it is seen that toward low solar activity the diurnal amplitude increases at low altitudes (300 km). The theory shows that this phenomenon follows from the lowering of the atomic oxygen layer during the decreasing phase of solar activity.

The pronounced increase (figure 4) of the scale height at higher altitudes (above 600 km) can be explained by the lowering of the helium layer. The smaller diurnal amplitude in the observational curves above 600 km indicates that a greater amount of helium must be present than was assumed in the theoretical models. Roemer [1963a,b] finds better agreement with data obtained from Echo 1 when using a theoretical model which has a 2.5 times greater amount of helium (I. Harris, unpublished calculations, 1963) than the theoretical models mentioned above. In figure 5 densities of this new model for $S=100$ are compared with the appropriate data by King-Hele. The agreement between theory and observations is obviously even better than in figure 4. In the models for higher levels of solar activities ($S\geq 150$), no noticeable increase of densities for heights up to 700 km occurs because of the increase of the number density N of helium by a factor of 2.5 at the boundary [$N(\mathrm{He}) = 6.25\cdot 10^{7}$ cm^{-3} at an altitude of 120 km]. For this reason, figures 2 and 3 are also applicable for the new models.

Also shown in figures 4 and 5 is a result obtained by Explorer 17 [Horowitz et al., 1963] which fills a gap in our knowledge at low altitudes (below 300 km) for times of low solar activity. At 2100 local time, April 3, 1963, at 260-km altitude and temperate latitudes, the measured density was $2.7\cdot 10^{-14}$ g cm^{-3}. The average solar 10.7-cm flux was about 75 in the usual units. Thus, if we use the relation shown in figure 1, a theoretical model with parameter $S=100$ is applicable.

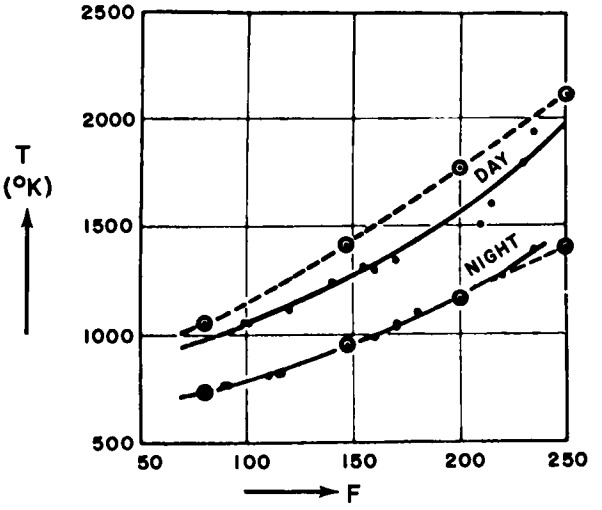


FIGURE 6.—Relation between exospheric temperature and the monthly averages of the solar 10.7-cm flux $\bar{F}$. The dots and small circles represent satellite drag data by Jacchia [1963] for nighttime and daytime, respectively. The temperatures are derived by means of Nicolet's model. The large circles give the temperatures of Harris–Priester models for 4 and 14 hours local time based on the empirical relation (figure 1).

If we plot the nighttime and daytime temperatures of the theoretical models using the new empirical relation (figure 1), we obtain an excellent agreement with the nighttime temperatures derived by Jacchia [1963] who used Nicolet's [1961] models for the conversion of observed densities into temperatures (figure 6). The daytime temperatures show a systematic difference of about 100° to 150°K. This could be explained by two reasons: (1) For a given density at any height Nicolet's models furnish one value for the temperature independently of local time, contrary to the Harris–Priester models, where the relation between density and temperature at a given height depends on local time (figure 7). This is because the Harris–Priester models are solutions of the time-dependent heat conduction equation. Therefore the use of Nicolet's models for conversion of density into temperature would lead to a diurnal temperature amplitude that is too small. The difference depends on the altitude and the level of solar activity. For the data used by Jacchia, the difference can be estimated to be about 50° to 100°K. (2) The theoretical models of Harris and Priester are based on the observational model of Bonn Observatory [Martin et al., 1961]. There are indications that the diurnal amplitude in this model is slightly too large, which again can account for a difference of 50° to 100°K.

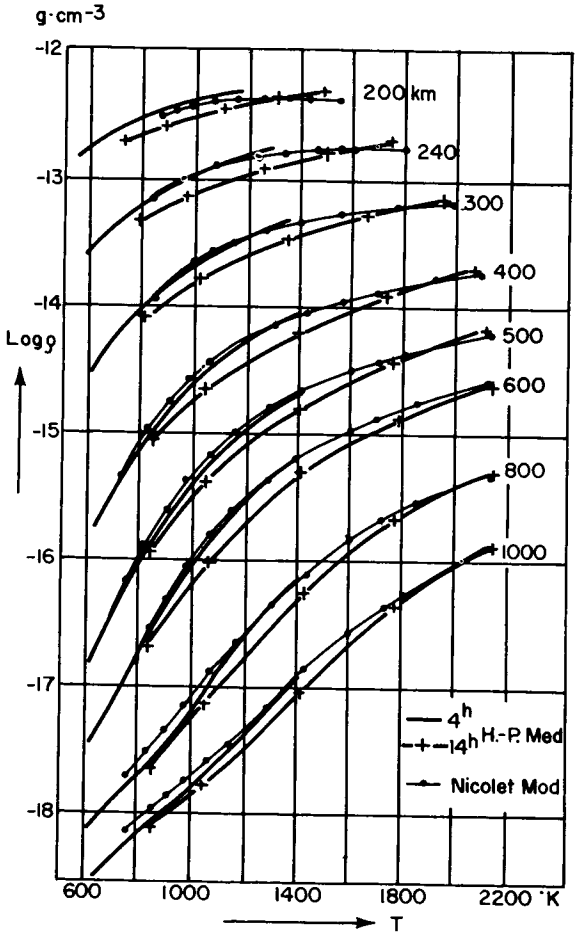


FIGURE 7.—Relation between density and temperature for eight different heights from 200 to 1000 km according to the Harris–Priester models for 0400 and 1400 local time (thick lines) and according to Nicolet's model (thin lines).

CONCLUSIONS

The comparison with air densities observed within the period from 1958 to 1963 has shown that the theoretical models give a good representation of the atmospheric properties and their changes during the decreasing phase of solar activity if the relation given in figure 1 is used. A still better agreement is obtained if in the theoretical models the amount of helium is increased by a factor of 2.5 at the boundary of 120 km. This, however, is important for periods of very low solar activity only. Further comparisons with forthcoming data for the years 1965 through 1968 will reveal whether the same empirical relation also holds for the increasing phase of the 11-year solar cycle.

REFERENCES

HARRIS, I., and W. PRIESTER, Theoretical models for the solar-cycle variation of the upper atmosphere, *J. Geophys. Res.* **67**, 4585–4591, 1962. (Also *NASA Tech. Note D*-1444, 261 pp., 1962.)

HOROWITZ, R., Report at Am. Geophys. Union Meeting, 44th, Washington, D. C., April 17–20, 1963.

JACCHIA, L. G., Atmospheric structure and its variations at heights above 200 km, report to COSPAR Working Group 4, presented at the Intern. Space Sci. Symp., 4th, Warsaw, June 1963.

JACCHIA, L. G., and J. SLOWEY, Accurate drag determinations for eight artificial satellites, in atmospheric densities and temperatures, *Smithsonian Astrophys. Obs. Spec. Rept.* 100, 117 pp., 1962.

KING-HELE, D. G., Decrease in upper-atmosphere density since the sunspot maximum of 1957–58, *Nature*, **198**, 832–834, 1963.

MARTIN, H. A., W. NEVELING, W. PRIESTER, and M. ROEMER, Model of the upper atmosphere from 130 through 1600 km derived from satellite orbits, in *Space Research, Proc. Intern. Space Sci. Symp., 2nd, Florence*, 1962, edited by H. C. van de Hulst, C. de Jager, and A. F. Moore, pp. 902–917, North-Holland Publishing Company, Amsterdam, 1961.

NICOLET, M., Density of the heterosphere related to temperature, *Smithsonian Astrophys. Obs. Spec. Rept.* 75, 30 pp., 1961.

PAETZOLD, H. K., Solar activity effects in the upper atmosphere deduced from satellite observations, in *Space Research, Proc. Intern. Space Sci. Symp., 3rd, Washington*, 1962, edited by W. Priester, pp. 28–52, North-Holland Publishing Company, Amsterdam, 1963.

ROEMER, M., Exospheric densities deduced from satellite drag data, paper presented at the Intern. Space Sci. Symp., 4th, Warsaw, June 1963.

ROEMER, M., Thesis, Bonn University, to be published in Veröff. der Univ.-Sternwarte Bonn, 1963.

INVESTIGATION OF PROBLEMS IN THERMAL CONVECTION

J. R. HERRING

Goddard Space Flight Center

The thermal convection equations for a thin layer of fluid are solved numerically as an initial value problem. The calculations include only those nonlinear terms which have the form of an interaction of a fluctuation in the velocity and temperature with the mean temperature field. In the present calculations, the velocity and temperature fluctuations have one horizontal wave number, and satisfy free boundary conditions on two conducting horizontal surfaces.

The computed steady state velocity and temperature amplitudes show many of the observed qualitative features. In particular, the experimentally observed boundary layering of the mean temperature field is correctly reproduced, and, at large Rayleigh number, the total heat transport is found to be proportional to the cube root of the Rayleigh number, provided the fluctuating temperature and velocity amplitudes have that horizontal wave number which maximizes the total heat transport. However, the heat transport found here for free boundaries is about three times the experimental value for rigid boundaries. The mean temperature gradient can become negative hear the boundaries for large Rayleigh numbers and large horizontal scale motions.

The linear stability of the system is also investigated, and it is concluded that the stable solutions for all Rayleigh numbers investigated ($R < 10^6$) have horizontal wave numbers which very nearly maximize the total heat transport. The stability study also indicates regions in which two or more horizontal wave numbers are required to support convection.

INTRODUCTION

This paper describes the results of a numerical investigation of the thermal convection equations for a thin layer of fluid confined between two plates on which free boundary conditions are employed. Our theoretical procedure is to include only those nonlinear terms which describe the interaction of the mean temperature with velocity and temperature fluctuations. That is, those terms responsible for eddy viscosity and eddy conductivity effects on the turbulence itself are omitted. The above eddy terms (hereafter referred to as fluctuating self-interactions) are discarded in a physically consistent manner, so that no unrelistic behavior results from their omission.

The motivation for this research is to discover quantitatively to what extent the turbulent convection problem can be comprehended without the fluctuating self-interactions. The system of equations obtained by deleting these terms corresponds to closing the hierarchy of moment equations at the first nontrivial level by discarding third order cumulants. The resulting system of equations is complete and involves no empirical parameters. Moreover, the gross energetics of the flow are preserved.

The method of numerical solution consists in integrating the Fourier amplitudes of the velocity and temperature fields forward in time until the steady state is achieved. This procedure has the advantage of assuring the stability of the final state provided a sufficient range of initial data is sampled. The present calculations, carried out on an IBM 7090 computer, contain one horizontal wave number and enough vertical wave numbers to ensure the elimination of truncation errors. In the numerical analysis, we have set the Prandtl number equal to unity. However, inspection of the equations which omit the fluctuating self-interactions shows that the heat transport is not

a function of Prandtl number, if the system is steady.*

The calculated velocity and temperature fields show many of the qualitative features of the experimentally determined fields. In particular, at large Rayleigh number, R, the total heat transport is found to be proportional to $R^{1/3}$, provided the velocity and temperature fluctuation fields have that horizontal wave number which maximizes the heat transport. However, the heat transport found here for free-boundary conditions is three times the experimental value for rigid-boundary conditions. Preliminary numerical studies of the rigid-boundary problem indicate that for large Rayleigh numbers ($R \sim 10^6$) the heat transport is about a factor of 2.3 smaller than that for free-boundaries and therefore approximately 50 percent higher than the experiment. Thus, it appears that the boundary conditions are quite important in producing the experimental heat transport. The system has two additional failings: it turns out that the fluctuating amplitudes are steady in time and the horizontal plan form of the motions is indeterminate.

The mean temperature gradient at low R closely resembles the experimental temperature profiles. At large Rayleigh numbers ($R \sim 10^6$), the gross features of the temperature profiles are correctly predicted by our system. The computed mean temperature gradients are large in a thin layer adjacent to the boundary and are quite small in the body of the fluid. The gradients near the boundary can become negative for motions of large horizontal scale, but remain positive for motions of a sufficiently small horizontal scale.

The stability of the steady state solutions against infinitesimal perturbation at other horizontal wave numbers is also investigated and the regions of instability are delineated. These results closely parallel perturbation results at low Rayleigh number and support the idea that the most stable solution is near the one transporting the most heat flux (Malkus and Veronis, 1058). A finite amplitude stability study, and the associated development of a several-horizontal-wave-number system to steady state will be the topic of a future investigation.

The idea that the important features of the convection problem are contained in the system which omits the fluctuating self-interactions is implicit in the theory of convection advanced by Malkus (1954). In the original formulation of his theory, Malkus sought a maximum for the heat transport subject to the constraint that the temperature gradient be positive, and that there be a smallest scale of motion participating in the advective heat transport. The smallest scale is supposed to be determined by the requirement that it be marginally stable in the presence of the mean temperature gradient occurring in the fluid. The smallest scale so determined furnished a cutoff in the cosine representation of the mean temperature gradient. The assumption that the heat transport was maximum under the above constraints then led to an explicit form for the temperature gradient.

A more recent formulation of the Malkus theory by Spiegel (1962) replaces the cosine representation of the temperature gradient by an expansion in terms of the set of eigenfunctions, which are marginally stable on the mean temperature gradient. This version of the Malkus theory is exactly equivalent to the system considered here, provided the horizontal scale of the motions is such that the mean temperature gradient is everywhere positive. In this sense, our numerical results contain, as a special case, the exact solutions to the Malkus theory for one horizontal wave number and free boundaries.

In this connection, a comparison of our computed temperature gradients with the predictions of the Malkus theory is relevant. In making this comparison, we must keep in mind that the system considered here is explicitly confined to only one horizontal wave number, whereas Malkus makes no explicit references to the nature of the horizontal-wave-number spectrum. We do not confirm the z^{-2} law for the gradient outside the boundary layer as predicted by Malkus, nor do we find a sharp cutoff in the cosine spectrum of the temperature gradients.

At low Rayleigh numbers ($R < 2000$) our numerical results are in agreement with the calculations of Malkus and Veronis (1958) and Kuo

* For a discussion of the Prandtl number dependence of the heat transport for the complete system, see Kraichnan (1962a).

(1961), who have obtained perturbation solutions to the convection equations. A procedure similar to ours has been used by Saltzman (1961) in studying the complete convection equations for $R \lesssim 6000$. Our approach differs from his in that we are able to allow very many vertical modes to be excited, whereas his results are limited to one vertical mode. Our results indicate that it is essential to allow many more vertical modes than horizontal modes, if large truncation errors are to be avoided. Thus, at $R=4000$, 10 vertical modes must be included to obtain realistic temperature profiles.

THEORY

The Equation of Motion and Boundary Conditions

We consider a thin layer of fluid confined between two infinitely conducting plates located at $z=0$ and $z=d$. The lower plate is maintained at zero degrees, and the top plate at a temperature $-T_0$, on an arbitrary temperature scale. The direction of gravity is specified by the unit vector $-\mathbf{k}$. The equations appropriate for our system are the Boussinesq approximations to the Navier–Stokes equations (Chandrasekhar, 1961, p. 16). We shall write these equations in a form in which the velocity and temperature ($\mathbf{v}$ and T) as well as the coordinate and time ($\mathbf{r}$ and t) are nondimensional. The only parameters entering the equations are then the Rayleigh number, R, and the Prandtl number, σ. The equations are

$$\nabla\cdot\mathbf{v}=0, \tag{1}$$

$$[\sigma^{-1}(\partial/\partial t)-\nabla^2]\nabla^2\mathbf{v}=\sigma^{-1}\nabla\times(\nabla\times(\mathbf{v}\cdot\nabla\mathbf{v}))+R\nabla\times(\nabla\times\mathbf{k}T), \tag{2}$$

$$[(\partial/\partial t)-\nabla^2]T=-\nabla\cdot(\mathbf{v}T). \tag{3}$$

Equation (2) is actually the double curl of the momentum equation, and hence the pressure variable is absent. The nondimensional variables are related to the dimensional ones (denoted by primes) in the following way:

$$\mathbf{v}=(d/\kappa)\mathbf{v}',$$
$$T=T'/d,$$
$$\mathbf{r}=\mathbf{r}'/d,$$
$$t=(\kappa/d^2)t'.$$

Here κ is the thermometric diffusivity of the fluid.

The boundary conditions on the velocity field are derived from the requirement that the fluid exert no shear on the confining plates. This, together with the continuity equation, implies that all even derivatives of the vertical velocity, w, vanish on the boundary. In the nondimensional notation the boundary conditions are

$$(\partial^m/\partial z^m)w(0,t)=(\partial^m/\partial z^m)w(1,t)=0, \qquad m=0,2,4,\cdots \tag{4}$$

and

$$T(0,t)=0; \qquad T(1,t)=-1. \tag{5}$$

Discard of the Fluctuating Self-Interactions

It is convenient to resolve the temperature field into a horizontal mean plus a fluctuating part;

$$T=-z+\psi(z,t)+\theta(\mathbf{r},t). \tag{6}$$

Here, $\psi(z,t)$ is a horizontally averaged distortion of the conduction state and $\theta(\mathbf{r},t)$ is the fluctuation of the temperature from its distorted value. In view of the boundary condition (5), and the interpretation of θ as a fluctuation from the horizontal mean, we may write

$$\psi(0,t)=\psi(1,t)=0 \tag{7}$$

$$\theta(x,y,0,t)=\theta(x,y,1,t)=\bar{\theta}=0. \tag{8}$$

The bar on (8) indicates an average over the horizontal. We now introduce (6) into (1), (2), and (3) and subtract from each of the resulting equations their respective horizontal mean. We find

$$[\sigma^{-1}(\partial/\partial t)-\nabla^2]\nabla^2 w=R\nabla_1^2\theta+\sigma^{-1}\{\nabla\times\nabla\times(\mathbf{v}\cdot\nabla\mathbf{v})\}_z,$$

$$[(\partial/\partial t)-\nabla^2]\theta=[1-(\partial\psi/\partial z)]w-\nabla\cdot(v\theta-\mathbf{k}\overline{w\theta}),$$

$$[(\partial/\partial t)-(\partial^2/\partial z^2)]\psi=-(\partial/\partial z)\overline{w\theta},$$

where

$$\nabla_1^2=(\partial^2/\partial x^2)+(\partial^2/\partial y^2).$$

There are two more equations, for the x- and y-components of the velocity field, but these are not necessary for our problem. The last terms in the equations above for w and θ have the form of a deviation of a bilinear fluctuating quantity from

its horizontal mean (fluctuating self-interaction). By discarding these terms we obtain the system to be investigated;

$$[\sigma^{-1}(\partial/\partial t) - \nabla^2]\nabla^2 w = R\nabla_1^2\theta, \tag{9}$$

$$[(\partial/\partial t) - \nabla^2]\theta = \beta w, \tag{10}$$

$$[(\partial/\partial t) - (\partial^2/\partial z^2)]\psi = (\partial/\partial z)\overline{w\theta}, \tag{11}$$

where

$$\beta(z) = 1 - (\partial\psi/\partial z) = -(\partial/\partial z)\bar{T}.$$

The significance of omitting fluctuating self-interaction can be expressed formally by examining the hierarchy of moment equations obtained from (1)–(3). By multiplying equations (2) and (3) by $v(t')$ and $T(t')$ and ensemble-averaging the appropriate sums of the resulting equations, we obtain the time evolution equations for the correlation coefficients $\langle v_i v_j' \rangle$, and $\langle v_i T' \rangle$, and $\langle TT' \rangle$. These equations couple the above second order moments to the transfer terms, which are cubic in $\mathbf{v}$ and T.

Since the system contains a non-vanishing first-order moment, ψ, the transfer terms contain both correlated third-order moments (cumulants) and products of first order moments with second-order moments. The discarding of the fluctuating self-interaction then corresponds to closing the system of moment equations by discarding the third order cumulants.† In the absence of mean fields this procedure would be empty.

We must now verify that our procedure of deleting third-order cumulants does not lead to physically unrealistic results. For our procedure to be acceptable, the system of (9), (10) and (11) must obey the conservation laws associated with the complete set of convection equations, and they must be free from unphysical consequences of the sort recently discussed by Ogura (1962), for a similar problem in isotropic turbulence. Ogura has demonstrated that the assumption of zero fourth-order cumulants and nonzero third-order cumulants is incompatible with a positive energy spectrum for all wave numbers.‡

† Discarding third order cumulants is quite different from discarding third order *moments*. The latter procedure has as a consequence that no steady state nontrivial amplitudes exist. For an investigation of the dynamics of decay for zero third-order moments see Deissler (1962).

‡ For a complete discussion of the cumulant discard approximations, see Kraichnan (1962a).

With regard to the last point, it should be noted that the positive definite character of the kinetic energy wave number spectrum and the spectrum for the square of the temperature field follows directly from the fact that it is possible to write the equations which delete third order cumulants in terms of amplitudes rather than moments. We observe that the amplitude (9), (10) and (11) all have real coefficients; hence, the square of any amplitude will remain positive for all time if the amplitude is initially a real number.

The conservation of entropy and kinetic energy are also preserved without the fluctuating self-interactions. By multiplying (10) by θ, (11) by ψ, and adding, we obtain after integrating over the entire volume of the fluid, the equation of conservation of entropy,

$$\tfrac{1}{2}(\partial/\partial t)\{|\theta|+|\psi|^2\}_v + \{|\nabla\psi|^2+|\nabla\theta|^2\}_v = \{w\theta\}_v. \tag{12}$$

Here the v subscript indicates an integration over the entire volume of the system.

We observe that equation (12), with a corresponding one for the conservation of the kinetic energy (Malkus and Veronis, 1958, p. 228) of the flow are exactly the same as those with the fluctuating self-interaction included. Contributions from the latter may be reduced to surface integrals which vanish.

Fourier Decomposition of the Equations

It is convenient to work with the Fourier components of (9)–(11) rather than their space-variable form. The free boundary conditions make the sine series appropriate. We therefore write:

$$w(\mathbf{r}, t) = \sum_{n,\alpha} f_\alpha(x, y) w_n^\alpha \sin n\pi z,$$

$$\theta(\mathbf{r}, t) = \sum_{n,\alpha} f_\alpha(x, y) \theta_n^\alpha \sin n\pi z,$$

$$\psi(z, t) = \sum_n \psi_n \sin n\pi z.$$

Here $f_\alpha(x, y)$ is an arbitrary set of orthonormal functions generated by the operator ∇_1^2, and obeying appropriate periodic boundary conditions

in the horizontal:

$$\nabla_1^2 f_\alpha(x, y) = -\pi^2\alpha^2 f_\alpha(x, y)$$

and

$$|\overline{f_\alpha f_{\alpha'}}|^2 = \delta_{\alpha\alpha'}.$$

Introducing the above representation into (9), (10) and (11) gives the following set of equations for the amplitudes w_n, θ_n, and ψ_n:

$$[\sigma^{-1}(\partial/\partial\tau) + n^2 + \alpha^2]\omega_n{}^\alpha = [\lambda\alpha^2/(n^2+\alpha^2)]\theta_n{}^\alpha, \quad (13)$$

$$[(\partial/\partial\tau) + n^2 + \alpha^2]\theta_n{}^\alpha = \omega_n{}^\alpha - \tfrac{1}{2}\pi \sum_{p=1} p\psi_p(\omega_{n+p}{}^\alpha + \sigma(n-p)\omega^\alpha{}_{|n-p|}), \quad (14)$$

$$[(\partial/\partial\tau) + n^2]\psi_n = \tfrac{1}{2}\pi n \sum_{p=1}\sum_\alpha \omega_p{}^\alpha(\theta_{n+p}{}^\alpha + \sigma(p-n)\theta^\alpha{}_{|n-p|}), \quad (15)$$

where

$$\lambda = R/\pi^4,$$

$$\tau = \pi^2 t,$$

$$\omega_n = w_n/\pi^2,$$

and

$$\sigma(x) = 1, \quad x>0$$
$$= 0, \quad x=0$$
$$= -1, \quad x<0.$$

Manipulation of the convolution terms in (14) and (15) is aided by the following identities:

$$\sum_p A_p(B_{n+p} + \sigma(n-p)B_{|n-p|}) = \sum_p B_p(A_{|n-p|} - A_{n+p}),$$

and

$$\sum_p A_p(B_{n+p} + \sigma(p-n)B_{|p-n|}) = \sum_p B_p(A_{n+p} + \sigma(p-n)A_{|p-n|}).$$

There are two conservation equations derivable from (14) and (15). The first is the Fourier representation of (12) for conservation of entropy. The other is the equation that partitions the total heat flux between conduction and convection; it is derived by multiplying (15) by $1/n$ and summing over n. We find

$$\sum_n [(1/n^2)(\partial/\partial t) + 1]\beta_n = \tfrac{1}{2}\pi^2 \sum_{n,\alpha} \omega_n{}^\alpha \theta_n{}^\alpha, \quad (16)$$

where

$$\beta_n = -\pi_n \psi_n.$$

Here the β_n's are the cosine transform of the mean temperature gradient. In the statistically steady state, (16) is the equation for the total heat flux, which is a constant of motion for the system. We now introduce a quantity $N(t)$, the total heat flux at the lower boundary:

$$N(t) = 1 + \sum_1^\infty \beta_n(t). \quad (17)$$

If the mean amplitudes are constant,

$$N = 1 + \tfrac{1}{2}\pi^2 \sum_{n,\alpha} \omega_n{}^\alpha \theta_n{}^\alpha.$$

In our units, the conduction state transports unit heat flux and this equation is the nondimensional form for the familiar equation for the total heat flux.

Structure of the Equation

Before proceeding to the numerical results, we give a brief resumé of the pertinent qualitative features of the system defined by (13), (14) and (15). First of all, we note that the horizontal wave numbers, α, are coupled only in their effect on the mean temperature field ψ. This interaction occurs diagonally in the sense that each α interacts only with itself. As a consequence there is a degeneracy in the horizontal plan form of the motion; the system is insensitive to the particular cell shape. Moreover, the number of α's is also indeterminant. The simplest situation is to have a single α support the motion and we investigate only this case here.

A single α will give nontrivial answers for the amplitudes w and θ only if it lies within a certain range. The range of α which will *not* support convection is obtained by assuming w and θ to be small, and demanding that they subsequently decay. If w and θ are small, ψ will be small to second order and our question is equivalent to that of marginal stability (See Chandrasekhar, 1961, p. 35). The system then will not support

convection if

$$(1+\alpha^2)^3/\alpha^2 \geq (R/\pi^4). \qquad (18)$$

Conversely, we assume that the steady state values of w and θ will be nonzero if α lies in the range complementary to (18).

The time behavior of the system is complicated by nonlinear effects. In the approach to the steady state, our numerical results indicate that the system executes overdamped oscillations with an ever increasing period of oscillation. This last remark is understandable since w and θ become marginally stable as $t > \infty$.

If the mean field, ψ, is statistically steady as $t \to \infty$, we may use a theorem of Spiegel (1962, p. 196) to show that w and θ are independent of time. Spiegel has shown that the principle of exchange of stability is valid (for free boundaries) in the presence of the mean gradient corresponding to the steady state solution to the mean temperature field given by (15). This implies that the growth rates for the appropriate eigen-function expansion for w and θ must all be zero in order for there to be a statistically steady state.

NUMERICAL PROCEDURE

In performing the numerical integration, we discard from the onset those Fourier amplitudes which will be zero in the steady state. We assume that the steady state amplitudes ω, θ and β have even parity about the mid-point $z = \frac{1}{2}$. This means that the even sine modes of ω and θ, and the odd cosine modes of β will have zero amplitude in the steady state. We therefore put their initial values equal to zero. The equations of motion (13), (14) and (15) then imply that the odd parity modes will remain zero for all subsequent time. Defining $\bar{\beta}_n = \beta_{2n}$, we may rewrite (14) and (15) in a more convenient form:

$$[(\partial/\partial\tau) + n^2 + \alpha^2]\theta_n = \omega_n + \tfrac{1}{2}\sum_{p=1}^{\infty} \bar{\beta}_p(\omega_{n+2p} + \sigma(n-2p)\omega_{|n-2p|}), \qquad (14')$$

$$[(\partial/\partial\tau) + 4n^2]\bar{\beta}_n = -2\pi^2 n^2 \sum_{p=1}^{\infty} \omega_p(\theta_{2n-p} + \sigma(p-2n)\theta_{|2n-p|}). \qquad (15')$$

Here, we have dropped the α superscript since we are interested in the system containing only one α. Equation (13) remains unchanged and the total heat flux is computed from (17).

Our procedure for integrating these equations is to assign an initial set of amplitudes to ω_n, θ_n, $\bar{\beta}_n$, and allow the system to evolve to the steady state. In doing so, we must truncate the infinite set of equations. Our procedure in this matter is to set all amplitudes ω_n, θ_n, $\bar{\beta}_n$ of index greater than a certain integer, n_0, equal to zero. This method of truncation guarantees exact conservation of heat flux and entropy for the abbreviated system. Since ω_1 is generally large and the θ_n's decrease rather slowly, we see from (15') that amplitudes for $\bar{\beta}_n$ above $n_0/2$ will have significant truncation errors. Truncation errors are assumed to be negligible if increasing n_0 does not appreciably alter the value of total heat transport. The total number of $\bar{\beta}_N$ modes included in these calculations ranged from 20 modes at $R = 4000$ to 80 modes at $R = 10^6$. The errors in the total heat transport due to the above are estimated to be less than one part in 10^3.

The integration forward in time was continued until constancy of heat flux (16) and entropy (12) was achieved to one part in 10^4. The time, in τ units necessary to achieve this ran from

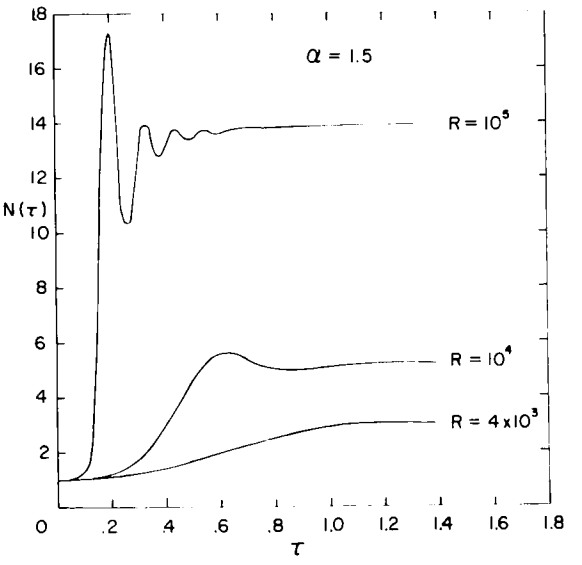


FIGURE 1.—Time development of the total heat flux, $N(\tau)$ for $R = 4\times10^3$, 10^4, 10^5 and $\alpha = 1.5$. The system is in the conduction state at $\tau = 0$, with all fluctuating amplitudes except w_1 equal to zero.

$\sim$1.4 at R$=4\times10^3$ to 0.3 at R$=10^6$. At high Rayleigh numbers, this criterion was not too satisfactory, since constancy of heat flux and entropy were achieved long before the amplitudes w and θ became steady. For these cases, it was necessary to check the time derivatives of the slowest evolving amplitudes, w_1 and θ_1. The system was observed to be steady if the derivative of w_1 was less than 2 percent of w_1.

Examples of the time evolution for the total heat transport $N(\tau)$ are given in Fig. 1, for R$=4\times10^3$, 10^4, 10^5 and $\alpha=1.5$. The system was started in the conduction state at $\tau=0$, with all fluctuating amplitudes ω_n and θ_n equal to zero,

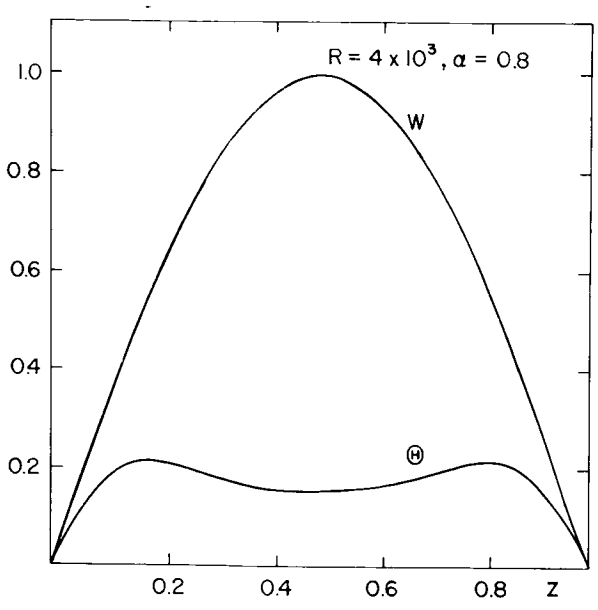


FIGURE 2.—$4.22\times10^{-2}w$ and 4.07θ for $R=4\times10^3$ and $\alpha=0.8$.

except ω_1, which has an initial value of unity. The convection is seen to develop initially by way of large oscillations, and to decay to the steady state with overdamped oscillations, whose period becomes increasingly larger. The time scale of the initial oscillations in these curves is of the order of the growth rate time in the conduction state.

DISCUSSION OF RESULTS

The computed steady state amplitudes are shown in Figs. 2–13. The normalization for β,

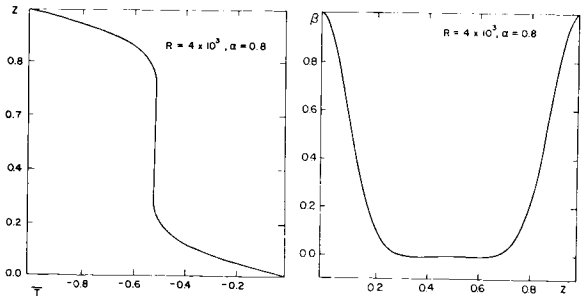


FIGURE 3.—(a) Mean temperature, $\bar{T}(z)$, for $R=4\times10^3$ and $\alpha=0.8$. (b) Mean gradient, $\beta(z)$, for $R=4\times10^3$ and $\alpha=0.8$. $\beta(z)$ is normalized by the total heat transport, $N=3.92$.

w and θ is given in the captions, while $\bar{T}(z)$ requires no normalization. The graphs of $\bar{T}(z)$ are in a reflected coordinate system to conform with an accepted procedure. The values of α in Figs. 2–9 were chosen so that the heat transport is very near its maximum. We now discuss in some detail the physical features of the steady state amplitudes ω, θ, β and $\bar{T}$.

The Mean Temperature $\bar{T}$, and mean gradient $\beta(z)$

The mean fields, $\bar{T}$ and β in Figs. 3, 5, 7, 9, 11 and 13 have an interesting behavior near the boundaries. At low Rayleigh number, these fields

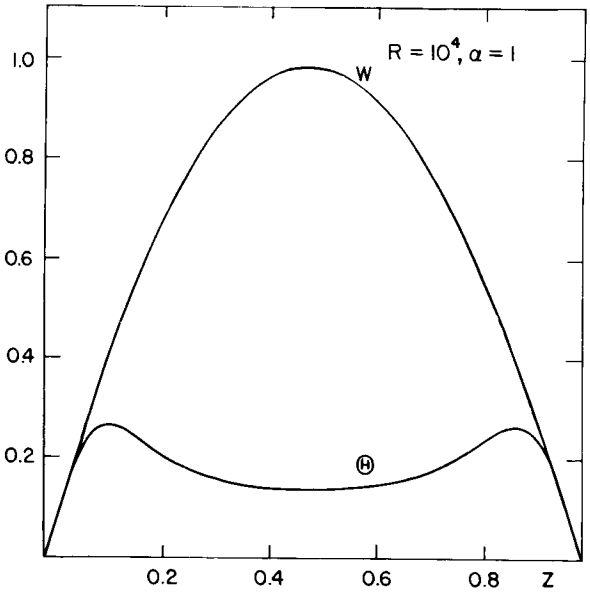


FIGURE 4.—$2.05\times10^{-2}w$ and 5.16θ for $R=10^4$ and $\alpha=1.0$.

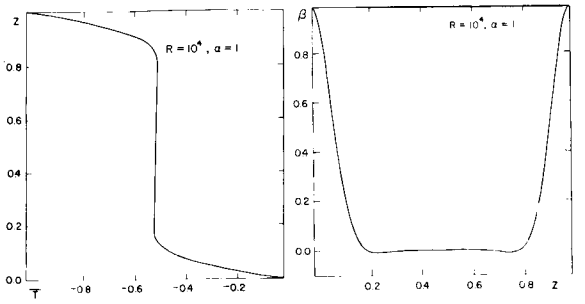


FIGURE 5.—Mean temperature, $\bar{T}(z)$, for $R=10^4$ and $\alpha=1.0$. (b) Mean gradient, $\beta(z)$, for $R=10^4$ and $\alpha=1.0$. $\beta(z)$ is normalized by the total heat transport, $N=5.82$.

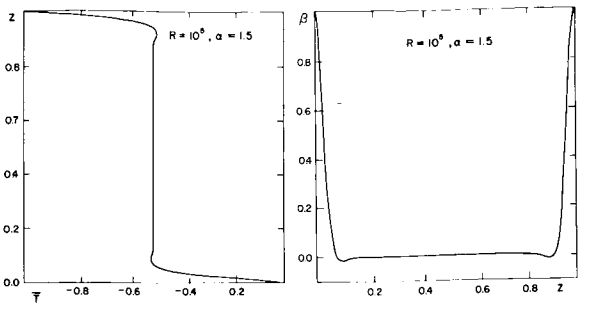


FIGURE 7.—(a) Mean temperature, $\bar{T}(z)$, for $R=10^5$ and $\alpha=1.5$. (b) Mean gradient, $\beta(z)$ for $R=10^5$ and $\alpha=1.5$. $\beta(z)$ is normalized by the total heat transport, $N=13.82$.

closely resemble the perturbation results of Markus and Veronis (1958), but the temperature gradient is slightly negative in the central region. As the Rayleigh number is increased, the negative temperature region collects closer to the boundary while in the central region, the temperature gradient becomes extremely small but positive.

The negative temperature gradient boundary region is apparently produced by an overshoot phenomenon. These occur typically for motions of large horizontal scale (small α) and disappear for motion of small horizontal scale. (See Figs. 9, 11 and 13). If the motion has a large horizontal scale, an element of fluid close to the lower plate moves in a region of high temperature for a relatively long time. When it eventually turns upward, it moves unchecked by eddy processes and penetrates the body of the fluid with an excessive heat flux. The mean gradient accommodates this motion by turning negative. The negative β region then checks the velocity field, so that the advective heat transport decreases toward the middle of the fluid. We see here evidence for the nonlocal property of the flow; if the Rayleigh criteria for convection were applicable locally, a negative β region would not persist in the steady state.

For motions of small horizontal scale (Figs. 11 and 13) the situation is somewhat different.

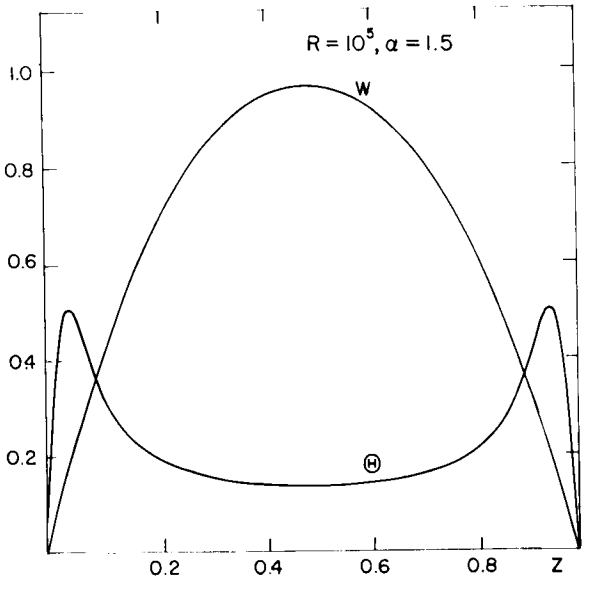


FIGURE 6.—4.33×10^{-3} and 9.42θ for $R=10^5$ and $\alpha=1.5$.

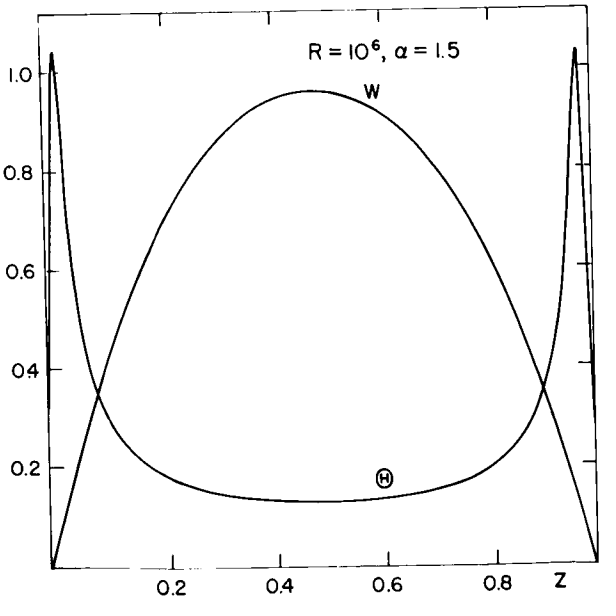


FIGURE 8.—$8.98\times10^{-4}w$ and 19.4θ for $R=10^6$ and $\alpha=1.5$.

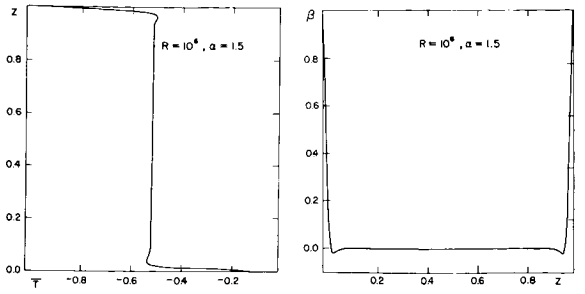


FIGURE 9.—(a) Mean temperature, $\bar{T}(z)$, for $R=10^5$ and $\alpha=1.5$. (b) Mean gradient, $\beta(z)$, for $R=10^5$ and $\alpha=1.5$. $\beta(z)$ is normalized by the total heat transport, $N=31.48$.

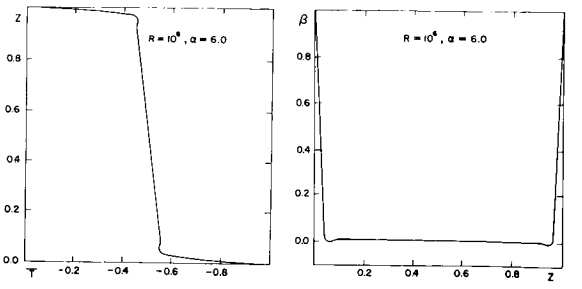


FIGURE 11.—(a) Mean temperature $\bar{T}(z)$, for $R=10^6$ and $\alpha=6.0$. (b) Mean gradient, $\beta(z)$, for $R=10^6$ and $\alpha=6.0$. $\beta(z)$ is normalized by the total heat transport, $N=22.3$.

In this case, an element absorbs little heat from the lower boundary region and loses it quickly by conduction because it belongs to a vertically elongated cell pattern. It also loses momentum by viscous drag, and attains its terminal velocity before reaching the central region of the fluid (see Figs. 10 and 12). To maintain constancy of heat flux the central region must conduct rather strongly, so that the mean gradient becomes large there.

Velocity and Temperature Fluctuations

The velocity and temperature fluctuation fields are shown in Figs. 2, 4, 6, 8, 10 and 12. We observe that the velocity fields, for all Rayleigh numbers, have an extremely large first mode. For example, at R = 4000 (Fig. 2) w_1 represents 99 percent of the total velocity amplitude, while at $R=10^6$ (Fig. 8) w_1 is 95 percent of the total. On the other hand, the θ_n modes decrease rather slowly as n increases.

The above behavior of the ω_n and θ_n spectra displays the character of the nonlinear coupling in our system. Thus, the term $\omega_1\theta_{2n-1}$ tends to be the dominant contributor to $\bar{\beta}_n$ [see equation (15′)] for reasonable small n. Conversely, terms of the form $\bar{\beta}_n\omega_1$ and $\bar{\beta}_{n-1}\omega_1$ tend to be the dominant contributors to θ_n [equation (14′)]. The nonlinear coupling scheme in the equations of motion is therefore highly nondiagonal, as opposed to the case of isotropic turbulence.

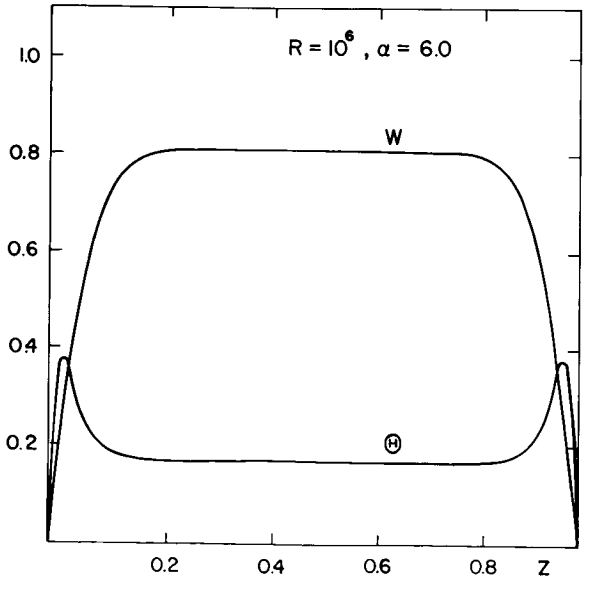


FIGURE 10.—$3.22\times10^{-3}w$ and 8.57θ for $R=10^6$ and $\alpha=6.0$.

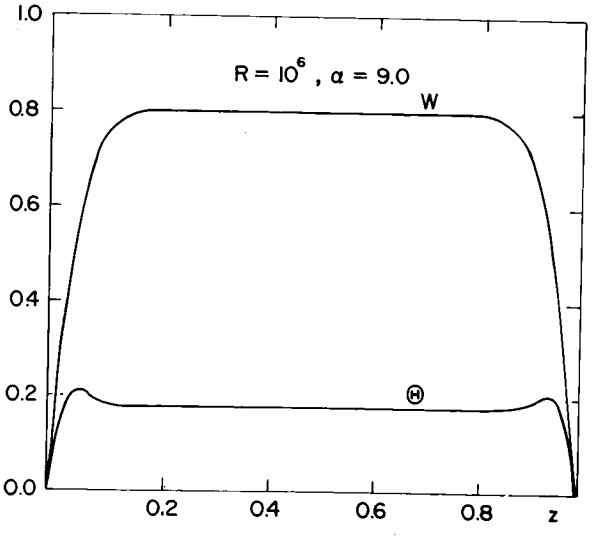


FIGURE 12.—$1.06\times10^{-2}w$ and 12.9θ for $R=10^6$ and $\alpha=9.0$.

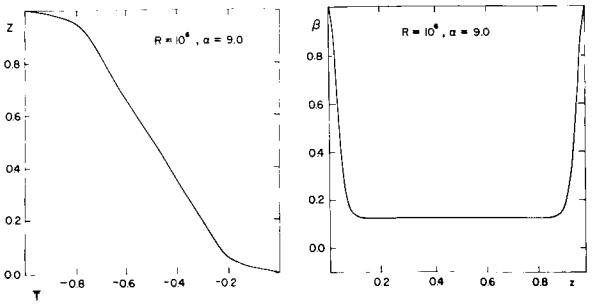


FIGURE 13.—(a) Mean temperature, $\bar{T}(z)$, for $R=10^6$ and $\alpha=9.0$. (b) Mean gradient, $\beta(z)$, for $R=10^6$ and $\alpha=9.0$. $\beta(z)$ is normalized by the total heat transport, $N=5.40$.

The strong nondiagonal coupling in the system of Fourier modes is a result of the distortion of the mean temperature profile combined with the pressure and dissipative forces for incompressible flow. The above forces are directly responsible for the occurrence of sixth-order derivatives in the marginal stability problem, of which the steady state amplitudes ω and θ are solutions in the presence of the mean field β. If we solve for the velocity modes $\omega_{n'}$, in the presence of the mean gradient β, by using the iteration technique of Section 5, we see that the higher modes of ω_n are suppressed by a factor $\sim n^{-6}$. For a reasonable β, this factor results in the higher ω_n modes making only a small contribution to ω.

Temperature Gradient Spectrum

The cosine spectrum of the mean temperature gradient, β, is given in Fig. 14 for $R=10^4$, 10^5, 10^6 and $\alpha=1.5$. We have connected the points with a smooth curve for the sake of clarity. We notice a tendency for the lower modes to saturate at $\beta_n=2$, which corresponds to the small gradient outside the boundary layer. In fact, if $\beta_n=2$ for all n, $\beta(z)$ is a δ function, and the gradient vanishes everywhere except at the boundary, where it becomes singular. At large Rayleigh numbers, the β_n spectrum is nearly Gaussian for small n, but decreases more rapidly at large n.

The tendency for the $\bar{\beta}_n$'s (for small n) to approach 2 as an upper bound is closely connected with the fact that the velocity field is marginally stable on the mean temperature gradient, β. This feature is brought out more clearly by examining the relation connecting the mean gradient field, $\bar{\beta}_n$ and the Rayleigh number R. Using the iteration method of Section V, we find

$$\frac{(1+\alpha^2)^3}{\alpha^2}\frac{\pi^4}{R}=1-\tfrac{1}{2}\beta_1+\frac{1/4}{1-\frac{1}{2}\beta_1}\sum_1^\infty\frac{(1+\alpha^2)^3}{[(2n+1)^2+\alpha^2]^3}\times(\bar{\beta}_n-\bar{\beta}_{n+1})^2+\cdots. \quad (19)$$

This series for R^{-1} converges rather rapidly for all the $\bar{\beta}_n$ which have been computed, and the terms explicitly written in (19) give R to an accuracy of $\lesssim 20$ percent at $R=10^6$. We note that for this equation to balance at large R, $\bar{\beta}_1$ must approach 2, and the remaining lower modes must decrease rather slowly as n increases.

The computed spectra (Fig. 14) are qualitatively quite different from the one derived by Markus (1954, p. 200). His spectrum is given by

$$\bar{\beta}_n=2\{1-[n/(2n_0+1)]\}.$$

Here $2n_0+1$ is a cut-off in the $\bar{\beta}_n$ spectrum, and it is the total heat flux in our units. Marginal stability is achieved at a much lower Rayleigh number for this spectrum than for the ones computed here.

With regard to the Malkus theory, Figs. 10 and 11 are relevant. For this case ($R=10^6$, $\alpha=6.0$) the temperature gradient is everywhere positive except near the boundaries where it approaches zero. The fields in Figs. 10 and 11 therefore fulfill all the requirements of the Malkus

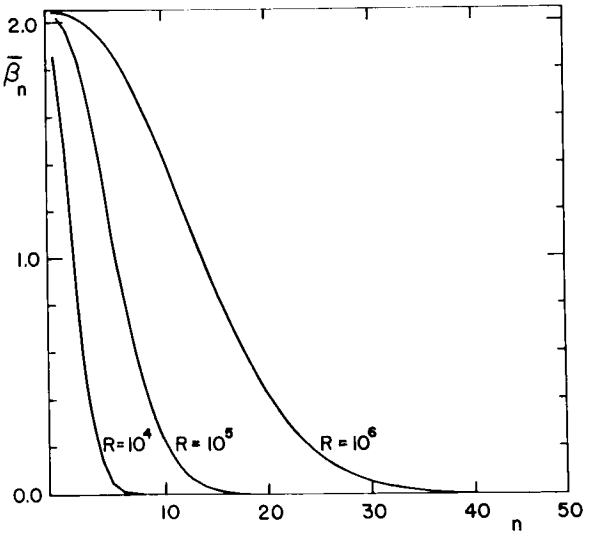


FIGURE 14.—Cosine spectrum of the mean temperature gradient for $R=10^4$, 10^5, 10^6 and $\alpha=1.5$.

theory as formulated by Spiegel (1962). We note for this case that the total heat flux is ~22, whereas Malkus obtains a heat transport of ~11 for free boundaries. In making this comparison, one should remember that these computations were made for a single horizontal wave number, whereas Malkus presumably allowed for a full spectrum of α's. However, if we interpret the computed heat transport as an upper bound to the heat transport as the Malkus theory prescribes, we conclude that for free boundaries the actual upper bound is at least a factor of two larger than that obtained by Malkus.

The Total Heat Transport as a Function of α and R

The total heat transport, as a function of R and α is given in Fig. 15. The Rayleigh numbers are indicated in the figure. These curves closely resemble the perturbation calculations at small R, but become increasingly broadened as the Rayleigh number increases. For a given R, the heat transport is entirely conductive ($N=1$) if α lies outside the bounds prescribed by (18). The value of α which maximizes the heat transport is $1/\sqrt{2}$ at the critical Rayleigh number (R=657), and apparently increases linearly in $R^{1/3}$, at large R. It is well represented at high Rayleigh numbers by the formula:

$$\alpha_{max} \simeq 0.7 + 0.01\ R^{1/3}. \qquad (20)$$

The data on this point is not entirely conclusive because of the large breadth of the curves. It

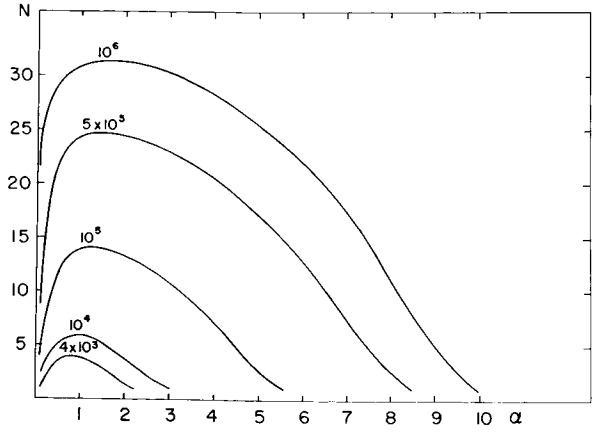


FIGURE 15.—The total heat transport N as a function of α for $R=4\times10^3$, 10^4, 10^5, 5×10^5 and 10^6.

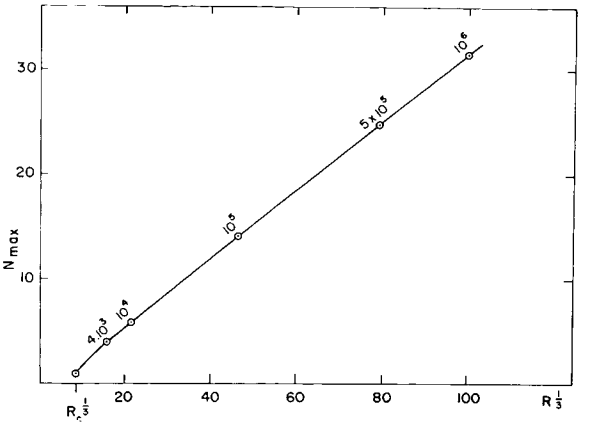


FIGURE 16.—Maximum total heat transport, N_{max} as a function of $R^{1/3}$

should be pointed out that (20) cannot be a correct asymptotic formula since α_{max} is proportional to $R^{1/3}$ and the value of α beyond which a single α cannot support convection is proportional to $R^{1/4}$ (equation 18).§ An estimate of the Rayleigh number beyond which (20) is incorrect is not warranted by the accuracy of the curves, but according to (18) and (20), it is $R \gtrsim 10^{18}$.

The maximum heat transport as a function of R is given in Fig. 16. For $R \geq 3000$ the data are accurately represented by the following $R^{1/3}$ law:

$$N \cong 0.31\ R^{1/3}. \qquad (21)$$

Experimentally, the Nusselt number N is ~0.085 $R^{1/3}$, for large R, and rigid boundaries (Jakob, 1959). We see no evidence for an intermediate $R^{1/4}$ law, but such a law may only be obtained in the rigid boundary problem. Below R~10^3, the data fits smoothly to the perturbation calculation of Malkus and Veronis (1958).

The discrepancy between (21) and the experiment is partly a result of eddy processes which our procedure omits. The use of rigid boundary conditions may improve the agreement, but if (21) is corrected for boundary effects as done by Malkus (1960) by decreasing N by $(657/1107)^{1/3}$, there remains a discrepancy of a factor of 2. If we choose horizontal wave numbers such that mean gradient is everywhere positive (Figs. 10 and 11) the discrepancy is reduced to 1.8. The

§ R. H. Kraichnan has pointed out to the writer that the heat transport, N_{max}, is then asymptotically proportional to $R^{3/10}$.

latter fields, however, have the unattractive feature of having a large temperature gradient in the central region of the fluid.

LINEAR STABILITY OF THE FIELDS

The velocity and temperature fields we have so far discussed are stable against the introduction of a disturbance of the *same* horizontal wave number for which the fields were computed. This stability is inherent in the method of integrating the equations. The stability of the steady state amplitudes against disturbances at wave numbers α' other than that α which supports the convection process has not yet been assured in our calculations. The question of stability of the solutions against disturbances of finite amplitudes leads directly back to the multi-α system of equations (13), (14) and (15). We should assume a whole spectrum of α's are initially excited, let them evolve to steady state, and repeat the calculation for an ensemble of initial conditions. We shall be content here with an investigation of linear stability of the system. This problem has some intrinsic interest, but our main purpose is to lay the framework for an investigation of the multi-α system.

It is convenient to pose the linear stability problem in terms of the Fourier amplitudes [equations (13), (14) and (15)]. We suppose that the system $w_n{}^{\alpha_1}$, $\theta_{n\alpha}{}^1$ and β_n have their steady state values, introduce disturbances $\delta\omega_n{}^\alpha$, $\delta\theta_n{}^\alpha$ and $\delta\beta_n$ and ask whether the latter grow or decay. Since α_1 and α are not coupled, $\delta\beta_n$ must decay initially. The problem then is reduced to determining the growth rates for $\delta w_n{}^\alpha$ and $\delta\theta_n{}^\alpha$ in the presence of the mean gradient β_n. Since exchange of stabilities has been proved for this system (Spiegel, 1962) we know that the system $w_n{}^{\alpha_1}$, $\theta_n{}^{\alpha_1}$, and β_n will be stable if the smallest critical Rayleigh number R_c for the perturbation system, $\delta\omega_n{}^\alpha$, $\delta\theta_n{}^\alpha$ is larger than the Rayleigh number for which w^{α_1} and θ^{α_1} were computed.

The marginally stable amplitude $\delta w_n{}^\alpha$, and $\delta\theta_n{}^\alpha$ satisfy (13) and (15) at a Rayleigh number R_c, with the time derivatives put equal to zero. Since the smallest R_c takes an eigen function even about $z=\frac{1}{2}$, we may abbreviate the perturbation system by eliminating the even sine modes from the velocity and temperature fluctuations. Defining

$$\varphi_n = \delta w_{2n-1},$$

$$\varphi_n = \delta w_{2n-1},$$

$$\bar{\beta}_n = \beta_{2n},$$

we may eliminate $\delta\theta_n$ by using the steady state form of (13) and write the marginally stability problem in the following matrix form:

$$A(\bar{\beta})\varphi = \mu\varphi, \tag{22}$$

where

$$A(\bar{\beta}) = \{\alpha^2/[(2n-1)^2+\alpha^2]^3\} \times \{\delta_{nm}+\tfrac{1}{2}(\bar{\beta}_{|n-m|}-\bar{\beta}_{n+m-1})\},$$

$$\mu = \pi^4/\mathrm{R}_c, \qquad \bar{\beta}_0 = 0.$$

In writing the matrix A, we have used the alternative form for the convolution term in (14).

The largest eigen value, μ (smallest R_c), of (22) may be obtained by the matrix iteration technique (Hildebrand, 1952). Since the first sine mode of the velocity will be largest, we may conveniently begin the iteration on a vector containing only this mode. Defining

$$|\,1\,\rangle = (1, 0, 0, \cdots, 0, \cdots)$$

we may write

$$[\pi^4/\mathrm{R}_c(\alpha)] = \mu_{\max} = \lim_{n\to\infty}(\langle 1 \mid A^n \mid 1\rangle / \langle 1 \mid A^{n-1} \mid 1\rangle). \tag{23}$$

The convergence of the iteration scheme is quite

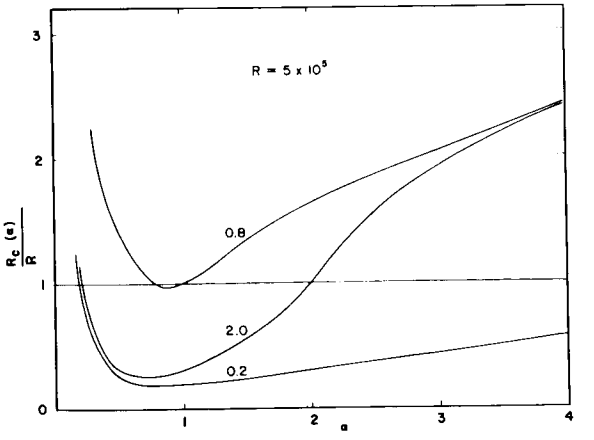


FIGURE 17.—Critical Rayleigh number R_c for $R=5\times10^5$ as a function of α. The value of α which supports the mean temperature fields labels the various curves.

rapid because of the structure of the A matrix. At the highest Rayleigh number considered, $R=10^6$, the 11th iteration gives R_c to one part in 10^6.

The calculated $R_c(\alpha)$'s are shown in Fig. 17 for $R=5\times10^5$. The particular α which supports the mean field, β, labels the various curves. At low Rayleigh numbers ($R<5.10^3$) these curves closely resemble those produced by perturbation calculations. Above $R\sim10^5$ they become increasingly distorted; steady state amplitudes of small horizontal wave numbers are enormously unstable with respect to an introduction of a disturbance at large α.

In Figs. 18 and 19 we give the zones of instability for the computed amplitudes for $R=10^4$ and $R=5\times10^5$. In these graphs, α_1 is the wave number that supports the convective process, and α_2 is the wave number of the perturbation amplitudes. The regions of instability are indicated by the shaded areas, whose outer boundaries are lines of marginal stability. The line $\alpha_1=\alpha_2$ is a trivial case of marginal stability. The value of α at which the two curves cross represents a solution which is infinitesimally stable against all other α's. This value of α begins at $1/\sqrt{2}$ at the critical Rayleigh number and increases slowly with increasing R. The rate of increase is seen to be slower than that α which maximizes the total heat transport. Referring to Fig. 15 we see that the use of the most stable α instead of α_{max} will not appreciably change the total heat transport.

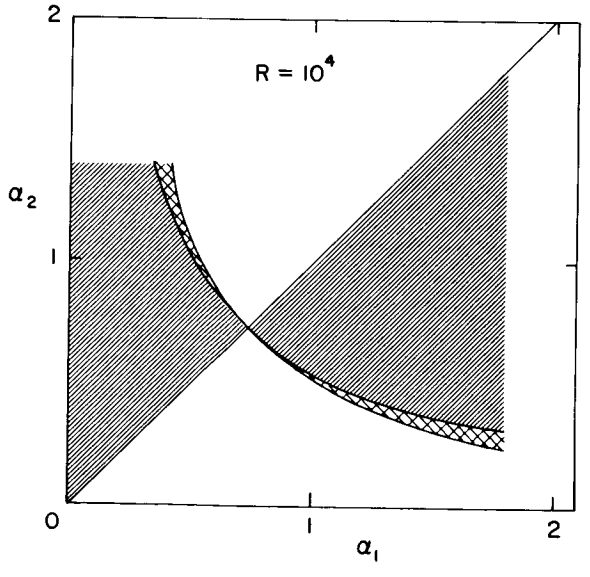


FIGURE 18.—Stability diagram for $R=10^4$. α_1 is the wave number that supports convection, and α_2 is the wave number at which a small perturbation is introduced. The shaded region indicates instability.

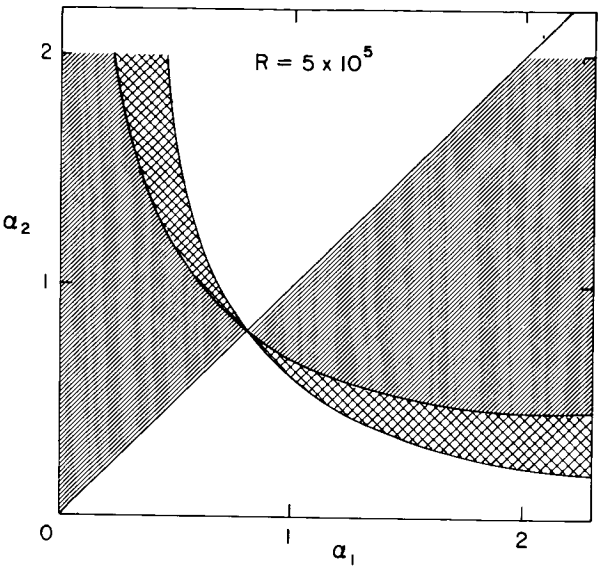


FIGURE 19.—Stability diagram for $R=5\times10^5$, α_1 is the wave number that supports convection, and α_2 is the wave number at which a small perturbation is introduced. The shaded regions indicate instability.

The zones which linear stability theory predicts must have two or more α's supporting convection are indicated by the cross-hatched regions shown in Figs. 18 and 19. These regions are obtained by perturbing the α_1 fields at α_2, assuming that the α_2 field subsequently dominates the convection, and then demanding that the α_2 fields be unstable with respect to a perturbation at α_1. The cross-hatched region is then bounded by the descending marginally stable curve and its reflection about the 45 deg line. At small Rayleigh number, $R<4\times10^3$, this are vanishes but it gradually increases with Rayleigh number.

CONCLUDING REMARKS

The temperature and velocity fields computed here with the fluctuating self-interactions absent show qualitatively reasonable behavior. The boundary-layering of the temperature field, which is found experimentally, is faithfully reproduced by the system, and the heat transport has the

experimentally determined dependence on Rayleigh number. In this respect, our results for the velocity and temperature amplitudes, as well as the stability analysis of the fields, confirm the original ideas of Malkus. However, our result for the heat transport for free boundary conditions does not agree quantitatively with Malkus.

The only disquieting features of the results are the negative temperature gradients which can occur near the boundary for small α, and the rather large amount of heat transported by the system. Aside from eddy processes, there are two other modifications in the system which must be explored before its quantitative accuracy can be properly accessed.

First, the use of the more realistic rigid boundary conditions will enable one to examine quantitatively the role of the eddy processes in producing the experimental temperature profile and the total heat flux. The presence of shear forces at the boundary will decrease the computed heat flux, and in checking the development of large scale horizontal motions there, it will reduce the negative temperature gradient. Preliminary indications are that the use of rigid-boundary conditions decrease the total heat transport by a factor of 2.3. Secondly, the introduction of several horizontal wave numbers will make the system more realistic, particularly at large Rayleigh numbers. It will also permit a study of finite amplitude stability of the system. The above modifications are currently under investigation and will be reported on in the near future.

ACKNOWLEDGMENTS

The author is very grateful to Dr. R. H. Kraichnan for suggesting this investigation, for many illuminating discussions through the completion of the work, and for making helpful suggestions regarding the presentation of material. He is also grateful to Dr. E. A. Spiegel for many valuable discussions, and for critically reading a final version of the manuscript.

It is a pleasure to acknowledge the able assistance of Mr. R. Rustin, who programmed and carried out the numerical computations on the IBM 7090 computer.

REFERENCES

CHANDRASEKHAR, S.: *Hydrodynamic and hydromagnetic stability*. Oxford, Clarendon Press, p. 16.

DEISSLER, R. G.: Turbulence in the presence of a vertical body force and temperature gradient. *J. geophys. Res.*, **67,** 3049–3062, 1962.

HILDEBRAND, F. B.: *Methods in applied mathematics*. Englewood Cliffs, N. J., Prentice-Hall, Inc., p. 66 *et. seq.*, 1952.

JAKOB, M.: *Heat transfer*. Vol. 1, New York, John Wiley and Sons, Inc., p. 535.

KRAICHNAN, R. H.: Turbulent thermal convection at arbitrary Prandtl Number. *Phys. Fluids*, **5,** 1374–1389, 1962a.

———: The closure problem of turbulence theory. *Proceedings of symposia in applied mathematics, Vol.* 13, *Hydrodynamic instability*, Providence, R. I., Amer. Math. Soc., 199–225, 1962b.

KUO, H. L.: Solution of the nonlinear equations of cellular convection and heat transport. *J. fluid Mech.*, **10,** 611–634, 1961.

MALKUS, W. V. R.: The heat transport and spectrum of thermal turbulence. *Proc. Roy. Soc. London* **A225,** 196–212, 1954.

———: Turbulence, notes of summer study program in geophysical fluid dynamics at the Woods Hole Oceanographic Institution. Unpublished, p. 38 *et seq.*, 1960.

———, and G. VERONIS: Finite amplitude cellular convection. *J. fluid Mech.*, **4,** 225–260, 1958.

OGURA, Y.: Energy transfer in an isotropic turbulent flow. *J. geophys. Res.*, **67,** 3143–3150, 1962.

SALTZMAN, B.: Finite amplitude free convection as an initial value problem. *J. atmos. Sci.*, **19,** 329–341, 1961.

SPIEGEL, E. A.: On the Malkus theory of turbulence. *Mechanique de la Turbulence*, Paris, C. N. R. S., 181–201, 1962.

HYDROGEN DIFFUSION IN THE THERMOSPHERE BY MONTE CARLO METHOD

S. O. KASTNER

Goddard Space Flight Center

The diffusion-limited flow of hydrogen in the thermosphere is obtained by a Monte Carlo calculation. The resulting density distribution has a steeper gradient than current theories predict, even taking omission of gravity into account. The departure is ascribed to increasing non-validity of the Chapman–Enskog diffusion coefficient as density changes over a mean free path become significant.

INTRODUCTION

The altitude distribution of minor nonpermanent constituents of the atmosphere and their rate of escape has been treated by several authors, among them Spitzer, (1952), Nicolet, (1960), Mange, (1961), and others. These authors in general base their analysis on the hypothesis that a minor constituent exists in hydrostatic equilibrium under its own partial pressure. Mange, (1961) and Bates and Patterson (1961) have taken flow into account to arrive at a steeper density gradient for hydrogen than would exist without flow, but hydrostatic equilibrium is still assumed. An earlier study which was not based on this hypothesis was carried out by Helge–Petersen (1928), who concluded that flowing helium did not attain hydrostatic equilibrium and therefore would decrease in density more rapidly with height than if such equilibrium were attained [see also Mitra (1952), p. 24].

Recently some indirect evidence has been obtained for the hypothesis that the hydrogen density gradient may be steeper than the current literature predicts. The intensity of night sky Lyman-α radiation was found in a rocket flight (J. M. Coogan, private communication) to decrease quite rapidly above 200 km, by 30% in 50 km. Such a decrease may reflect a steep density gradient, though the radiative transfer process is a complicated one.

Donahue and Thomas (1963), also, suggest that the Lyman-α radiation appears in part to originate by single scattering from an extensive hydrogen envelope which is denser than expected on the basis of current estimates of the escaping flux.

This motivated the present work in which the Monte Carlo method is applied to the problem of hydrogen diffusion in the thermosphere, the object being to avoid hypotheses such as that of hydrostatic equilibrium and to simulate actual collisions as closely as possible.

The physical model dealt with here is probably best described by a comparison with the analytic treatments of Nicolet (1960) and others. In such treatments the vertical transport (flow) velocity of a minor constituent of the atmosphere is given by

$$w = -D[n^{-1}(dn_1/dz) + (m_1 g/kT) + (1+\alpha_T)(dT/dz)] \quad (1)$$

where D is the diffusion coefficient, n_1 the number density of minor constituent, m_1 the mass of minor constituent, and α_T the thermal diffusion coefficient.

The first term on the RHS, represents usual diffusion due to concentration gradient. It is the major term and is the one essentially calculated here. The second and third represent the effect of gravity, while the fourth represents thermal diffusion; the latter is a relatively minor term.

In the present calculation we do not include

the effect of gravity, i.e., between collisions no force is assumed to act on the H atoms. The flow is therefore diffusion-limited only. The effect of this admittedly severe simplification will be discussed below.

The thermal diffusion term is believed to be automatically present in the calculation, since it results from exchange of momenta between the H atoms and N_2 molecules during elastic collisions.

APPLICATION OF MONTE CARLO METHOD

We assume for simplicity that the major atmospheric constituent is nitrogen, whose density and temperature depend on altitude as in the Cospar Atmosphere. The region dealt with lies between 100 km, and 500 km; this region is divided into 100 zones, each 4 km thick, each zone having the appropriate Cospar density and temperature.

Hydrogen atoms are introduced at levels of interest and made to undergo collisions with successive nitrogen molecules produced from appropriate Maxwell–Boltzmann (MB) distributions. In each collision, scattering is assumed to be isotropic in the center-of-mass system. The cross-section for H–N_2 collisions is not available either experimentally or theoretically, so that one must make an assumption concerning the interaction potential. A Lennard-Jones (6—12) potential was assumed here, as discussed below. Conveniently, the possibility of formation of an H–N_2 complex may be ruled out because of the high stability of N_2.

In somewhat more detail, the process constructed is as follows. Each zone z has an MB distribution of N_2 molecules determined by its temperature T. A hydrogen atom is injected into zone $z=1$, with velocity chosen at random from the MB distribution of this zone. An N_2 molecule is produced with velocity also chosen at random from the MB distribution. The collision cross-section, which depends on the relative velocities, is computed and leads to an actual path traversed. The collision then occurs, and the resulting velocity of the H atom is obtained. A new N_2 molecule is produced, and the process is repeated. Transitions from one zone to the next are handled much as in reference 8. However, in the upper zones, which have low density, a collision-free path may pass through many zones, so that a procedure was incorporated to add traversal times to the intervening zones.

THE HYDROGEN–NITROGEN COLLISION PROCESS

(a) As mentioned above, scattering is assumed to be isotropic in the center-of-mass system. The collision formulas of Cashwell and Everett, (1959) pp. 56–60, were used in their general form, i.e., as applied to collisions with a moving target particle rather than a fixed target.

(b) A cross-section for collisions between H atoms and N_2 molecules was obtained by means of "combining "rules", which are discussed by Hirschfelder, Curtis and Bird (1954) (hereafter HCB), p. 567. The intermolecular potential assumed was the Lennard-Jones "6–12" potential.

$$\phi(r) = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^{6}] \qquad (2)$$

where σ is the distance at which potential is zero and ϵ the depth of potential minimum. For two dissimilar molecules, the parameters σ and ϵ according to the combining rules are to be taken as

$$\sigma_{12} = \tfrac{1}{2}(\sigma_1 + \sigma_2)$$

$$\epsilon_{12} = (\epsilon_1 \epsilon_2)^{1/2}$$

where σ_1, ϵ_1, σ_2, ϵ_2 are the parameters which hold for like-molecular collisions of molecules 1 and 2, respectively.

The collision cross section for this intermolecular potential is (in the notation of HCB):

$$Q^{(1)} = \pi \sigma_{12}^2 Q^{(1)*}$$

where the reduced cross section $Q^{(1)*}$ is a function only of the reduced kinetic energy of relative motion g^{*2}:

$$g^{*2} = \tfrac{1}{2}\mu g^2/\epsilon_{12}$$

g being the initial relative speed of the colliding molecules and μ the reduced mass.

The quantity $Q^{(1)*}$ is plotted on page 558 of HCB: values were taken directly off this curve for the purpose of the calculation.

The parameters $\sigma_1 \equiv \sigma_H$, $\epsilon_1 \equiv \epsilon_H$, $\sigma_2 \equiv \sigma_{N_2}$, $\epsilon_2 \equiv \epsilon_{N_2}$ are thus needed. σ_{N2} and ϵ_{N_2} are given in table I–A of HCB., p. 1111, as 3.75×10^{-8} cm and

1.10×10^{-14} erg, respectively. For hydrogen, the Lennard-Jones parameters are not given, but the H_2 potential is well known and is given for example by Herzberg (1950). On the assumption that this actual potential closely resembles a Lennard-Jones potential, the effective values of σ_H and ϵ_H are 0.39×10^{-8} cm and 7.16×10^{-12} erg, respectively. The latter value is large because H_2 is a tightly bound molecule; this means that $\epsilon_{12}=(\epsilon_{N_2}\cdot\epsilon_H)^{1/2}$ is correspondingly large, implying a very large cross section, indeed, for small relative velocities. This is felt to be reasonable, however, and in any case the above combining rules seem to be the only ones available for obtaining intermolecular potentials for such dissimilar molecules.

(c) The effect of ignoring atomic oxygen may be briefly discussed at this point. The cross section for H–O collisions must be large, since the OH radical can be formed in three-body collisions. Thus the presence of O will hinder diffusion of H, decreasing the density gradient somewhat. It would be feasible to actually include a given proportion of O atom in the Monte Carlo calculation, to verify this qualitative statement, though this was not done here. Bates and Nicolet (1950) estimate that the density of O may be as high as 10^{12} atoms/cc at 95 km, with the O/N_2 ratio increasing with altitude, so that such a refinement might be worthwhile.

Formation and dissociation of OH can probably be neglected above 100 km, since according to Bates and Nicolet (1950) its density is only 10^4 atoms/cc at 90 km and drops rapidly.

CALCULATION OF TRANSMISSIONS AND FLUXES

If the density distribution of particles diffusing through an extended dense inhomogeneous medium is desired, the Monte Carlo method can be used but runs into the difficulty that insufficient numbers of particles penetrate the full extent of the medium. Kahn (1956) and others have devised methods of weighting ("splitting", etc.) to circumvent this difficulty.

An alternative method is used here which describes the diffusion problem in terms of a transmission and reflection problem, making use of the fact that transmission coefficients of finite

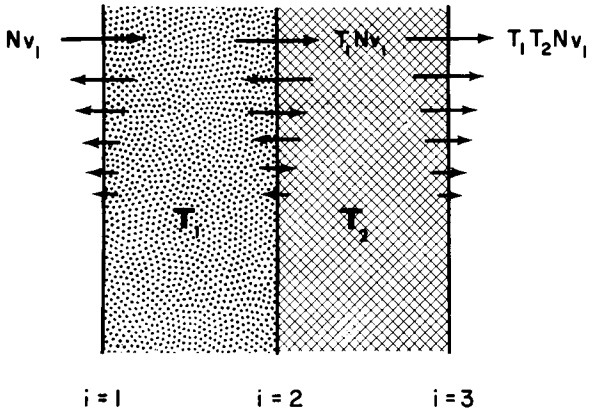


FIGURE 1.—Steady state fluxes in two regions.

regions of the medium are easily obtainable by the Monte Carlo method. The transmission relations used have been essentially derived (in more complex form) in such fields as multilayer optics (Heavens, 1955).

Consider two "slab" media in contact, whose transmission coefficients separately are T_1, T_2; i.e., out of a flux of Nv_1 particles/cm^2/sec injected into the first slab, a flux T_1Nv_1 will be transmitted. Of this, in turn, a fraction T_2 will be further transmitted through the second slab.

As shown in figure 1 there are multiply reflected and transmitted fluxes, and if the system is in a steady state (as much flux leaving the system as entering) one can add up all these partial fluxes, finding that:

(a) The transmitted flux at plane $i=3$ is given by

$$Nv_1T_1T_2/(1-R_1R_2) \qquad (R_1=1-T_1;\ R_2=1-T_2)$$

so that the over-all transmission is given by $T_1T_2/(1-R_1R_2)$.

(b) The forward flux at plane $i=2$ is

$$Nv_1T_1/(1-R_1R_2)$$

The backward flux at plane $i=2$ is

$$Nv_1T_1T_2/(1-R_1R_2)$$

so that the flux difference at this plane is

$$Nv_1T_1T_2/(1-R_1R_2)$$

in agreement with equation (1).

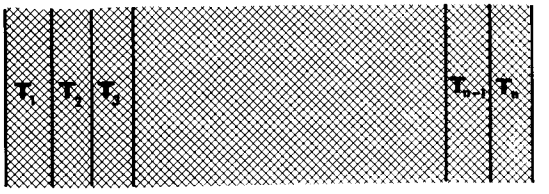


FIGURE 2.—Medium consisting of n regions.

Now consider a slab medium of thickness l in the z direction. Divide it by planes into n regions, as in figure 2, with the planes labelled by the index i. The transmission coefficients T_1, T_2, $\cdots$, T_n of the regions are assumed to be known (by injecting particles into one such region at a time).

The combined transmission of the last two regions is

$$T_1 = T_{n-1}T_n/(1-R_{n-1}R_n)$$

The combined transmission of the last three regions is then

$$T_2 = T_{n-2}T_1/(1-R_{n-2}R_1) \qquad (R_1 = 1-T_1)$$

By continuing this iteration process the over-all transmission is obtained:

$$T \equiv T_{n-1}$$

The forward and backward fluxes at any given plane i can also be obtained by computing the combined transmissions of the regions to the left of the plane and those to the right of the plane, and applying the two-region formulas above.

Let these fluxes at the plane i be denoted by $n_i^+\bar{v}_i$ and $n_i^-\bar{v}_i$, respectively. The total density at the plane i is $(n_i^+ + n_i^-)$; thus the total flux $\bar{v}_i(n_i^+ + n_i^-)$ must be divided by the mean particle velocity $_i$, which is not obtainable from the transmission measurements alone. To find $\bar{v}_i$ one must relate the transmission description of the system to its description as a diffusing process.

A given atom, moving in a random walk, undergoes a displacement ξ along the z axis in a time Δt. ξ^2, the mean square of such displacements for all atoms originating in a volume ΔV, is related (Kennard, 1938) to the diffusion coefficient D_i by

$$D_i = \lim_{\Delta t \to 0}(\bar{\xi}^2/\Delta t) \qquad (3)$$

The mean displacement $\bar{\xi}_i$ at the plane i also gives the velocity $\bar{v}_i$ required above, i.e.,

$$\bar{v}_i = \lim_{\Delta t \to 0}(\bar{\xi}_i/\Delta t) \qquad (4)$$

The total density at plane i is then given by

$$n_i^+ + n_i^- = \frac{(n_i^+ + n_i^-)\bar{v}_i}{\bar{v}_i} = \frac{f_i(N\bar{v}_i, T_1, T_2, \cdots, T_n)}{\bar{v}_i} \qquad (5)$$

where f_i is constructed as described above (combining transmissions to right and to left of the given plane, etc., to arrive at the sum of forward and backward fluxes).

The mean time $\bar{t}_i = 1/\bar{v}_i$ spent per atom in a unit volume at plane i is more directly gotten in practice than $\bar{v}_i$, so that (5) is used in the form

$$n_i = \bar{t}_i(n_i^+ + n_i^-)\bar{v}_i \qquad (5a)$$

In the actual computer program, atoms injected at each plane are always picked from an isotropic and Maxwellian distribution, a distribution which cannot strictly hold in the actual diffusion process. The error resulting from this approximation should be small, however.

RESULTS

Transmissions and Fluxes

Several thousand H atoms were successively injected at altitudes of 140, 152, 168, 184, 200, 296, 496 km, corresponding respectively to planes $i = 1, 2, 3, 4, 5, 6, 7$. The transmissions resulting are given in table 1.

From these, the values shown in table 2 were found for forward and backward fluxes, and sums and differences of these fluxes.

The over-all transmission of the atmosphere

TABLE 1

Region (km)	Transmission
140–152	0.0040±0.0015
152–168	0.0074±0.0026
168–184	0.0116±0.0034
184–200	0.0183±0.0037
200–296	0.0095±0.0037
296–496	0.0231±0.0051

TABLE 2

Altitude (km)	i	F_i^+	F_i^-	$S_i \equiv (F_i^+ + F_i^-)$	$M_i \equiv (F_i^+ - F_i^-)$
140	1	1.0000	0.9985	1.9985	0.0015
152	2	0.6281	0.6266	1.255	0.0015
168	3	0.4278	0.4263	0.854	0.0015
184	4	0.3005	0.2990	0.600	0.0015
200	5	0.2204	0.2189	0.439	0.0015
296	6	0.06466	0.06316	0.128	0.0015
496	7	0.0015	0.0	0.0015	0.0015

from 140 km to 500 km is therefore 0.0015, which is also the net upward flux when unit flux is injected at 140 km. (Extra significant figures included in the table entries are retained for calculational purposes, and are not intended to represent unwarranted precision.)

Diffusion Times and Relative Densities

At each of the planes i, H atoms were successively injected and timed as they diffused through a 4 km distance. The mean times and resulting relative densities are shown in table 3.

It is to be noted that $t_7 > t_6$. This appears to be a real effect, explainable physically as due to freedom of the atoms to travel horizontally for long distances without collisions.

These relative densities are plotted in figure 3, together with the main nitrogen atmosphere and the Bates–Patterson hydrogen atmosphere (Bates–Patterson, 1961). The latter, constructed for an exosphere temperature of 1500°K, is representative of the hydrogen distributions obtained by use of equation (1). The exponent k of the density scale is not specified because the absolute densities are not necessary here; k takes a different value for each curve.

TABLE 3

Altitude (km)	i	time t_i to diffuse 4 km (sec)	Relative density $t_i S_i$
140	1	68.1	136
152	2	25.5	32
168	3	16.4	14
184	4	11.5	6.9
200	5	9.0	3.95
296	6	2.1	0.27
496	7	4.4	0.0066

DISCUSSION OF RESULTS

Applicability of Net Flow Relation

The density curve obtained here has a much steeper slope than that of Bates and Patterson. One then asks whether the net flow relation

$$f = D_i (dn(\mathrm{H})/dz)_i \qquad (6)$$

is satisfied for this curve, where D_i is the diffusion coefficient at plane i, computed from the applic-

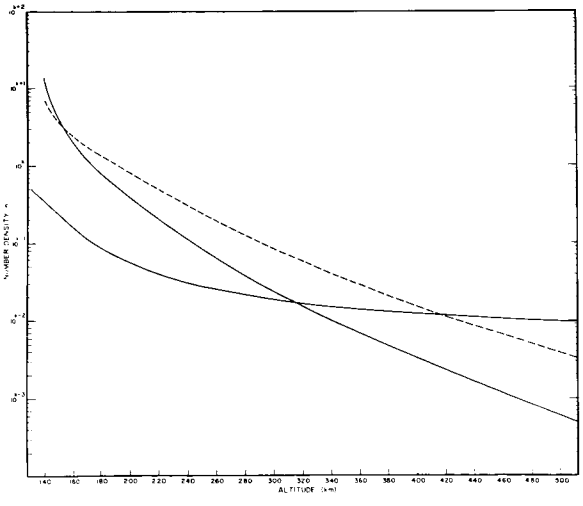


FIGURE 3.—Density distributions of assumed main nitrogen atmosphere (dashed), Bates-Patterson hydrogen atmosphere (dash-dot), hydrogen atmosphere derived here (solid); k is a different (unspecified) parameter for each curve.

TABLE 4

i	$(dn(\mathrm{H})/dz)_i$ (No./cm^4)	D_i (cm^2/sec)	$D_i(dn(\mathrm{H})/dz)_i$
1	48.7	1.54×10^{9}	7.5×10^{10}
2	17.8	3.11×10^{9}	5.5×10^{10}
3	6.54	5.35×10^{9}	3.5×10^{10}
4	2.8	8.22×10^{9}	2.3×10^{10}
5	1.3	1.26×10^{10}	1.65×10^{10}
6	0.059	1.10×10^{11}	0.65×10^{10}
7	0.0011	2.40×10^{12}	0.26×10^{10}

able formula on page 539 of HCB:

$$D_{12}=0.00268\frac{[T^3(M_i+M_2)/M_1M_2]^{1/2}}{p\sigma_{12}^2\Omega_{12}^{(1,1)*}T_{12}^*} \qquad (7)$$

where p = pressure in atmospheres; $T_{12}^*=kT/\epsilon_{12}$; σ_{12}, ϵ_{12}/k are potential parameters in Å and °K, respectively; $\Omega_{12}^{(1,1)*}$ is an integral given in table I–M, page 1126 of HCB's appendix, as a function of T.

The product

$$D_i\cdot(dn\ (\mathrm{H})/dz)_i$$

is given in table 4. The product is seen to decrease more and more rapidly with altitude, instead of remaining constant as one would normally expect.

The explanation for this nonconstancy is believed to be the fact that the Chapman–Enskog kinetic theory, from which the diffusion coefficient is derived, becomes invalid when density and temperature changes are appreciable over a mean free path (HCB, pp. 18–20). An effective diffusion coefficient is such a case does not appear to have been derived in the literature. The situation has been dealt with to some extent in the field of gas dynamics (Schaaf and Chambre, 1961), which we may qualitatively discuss to show its bearing on the atmospheric problem.

The three main flow regimes of gas dynamics are the continuum flow regime, the free molecular flow regime, and the transition regime lying between these. The regimes are based on the range of values of a Knudsen number K which is defined as the ratio of mean free path to some significant dimension of the flow field. Very roughly, the continuum regime corresponds to $K\ll1$, the free molecule regime to $K\geq1$, and the transition regime to intermediate values of K. Though this division of gas dynamics pertains actually to flow past solid surfaces, it appears relevant in the present case to consider density scale height as a significant dimension. Knudsen numbers $K_i=\bar{l}_i/h$ are listed in table 5, using mean free paths $\bar{l}_i$ taken from the computer calculation. K becomes appreciable compared to unity as altitude increases, so that conditions depart from those of the continuum regime and become those of the transition regime. The diffusion coefficient of the continuum regime will not then be valid.

The above does not consitute a proof of the suggested explanation, of course, but only an argument for its plausibility.

A direct measure of the departure of the diffusion coefficient formula from validity is afforded by the entries of table 4; since the product $D_i\cdot(dn\ (\mathrm{H})/dz)_i$ decreases by a factor of 11.5 between the altitudes of 140 km and 296 km, D at 296 km is too small by this factor.

Effects of Approximations

Two important approximations have been made in arriving at this density curve which could be examined at this point. As noted earlier, gravity has been neglected. Its inclusion would decrease net flow because molecular escape from the critical level near 500 km would then be an important factor in controlling the over-all escape rate. The relative importance of molecularly limited escape and diffusion-limited escape is reviewed by Urey (1959). At lower temperatures (~1500°K) of the

TABLE 5

i	Mean free path $\bar{l}_i$ (km)	Scale height h_i (km)	$K_i=\bar{l}_i/h_i$
1	0.036	25	0.0014
2	0.10	33	0.0030
3	0.12	38	0.0031
4	0.21	39	0.0054
5	0.35	41	0.0085
6	2.0	53	0.038
7	25.5	83	0.31

critical level, the former predominates so that the curve obtained in the present work, which assumes completely diffusion-limited escape, is somewhat too steep (net flow being too high gives an increased density gradient). Correcting for gravity would not however bring the curve up to the values of Bates and Patterson, for example, since their values were obtained using the usual diffusion coefficient.

A lesser approximation was made in setting density of the main atmosphere equal to zero above 500 km. Collisions above this altitude would probably have a negligible effect on the diffusion process.

Noninclusion of Hydrostatic Equilibrium

The assumption, adopted at the start, that the diffusing hydrogen is not in hydrostatic equilibrium appears to play a lesser role in steepening the density curve than does the nonvalidity of the diffusion coefficient, though it is difficult to separate out its effect on the final result.

CONCLUSIONS

The Monte Carlo calculation of diffusion of hydrogen through a nitrogen atmosphere results in a higher gradient of hydrogen density than current theories predict, even allowing for the fact that gravity was not included. The departure is ascribed to increasing nonvalidity of the usual diffusion coefficient with altitude, because the mean free path becomes significantly large compared to density scale height. The usual (calculated) diffusion coefficient is estimated to be an order of magnitude too low at 300 km altitude.

The density curve obtained is somewhat too steep, because of the omission of gravity. The true curve will have a gradient which is less steep though steeper than hydrogen distributions calculated on the basis of the usual diffusion coefficient.

Since the diffusion-limited flow is found to be higher than in other calculations, the over-all escape flux of hydrogen will also be higher. This could account for the denser hydrogen envelope hypothesized by Donahue and Thomas.

ACKNOWLEDGMENT

The author thanks J. E. Kupperian for suggesting the problem and for valuable discussions in the course of the work. C. Wade programmed the calculation and made the many necessary runs. Talks with C. L. Wolff were appreciated also.

REFERENCES

BATES, D. R., and M. NICOLET, *Jour. Geophys. Res.*, **55,** 301, (1950).

BATES, D. R., and T. N. L. PATTERSON, *Planet. Space Sci.*, **5,** (1961), 257; ibid 5, 328, (1961).

CASHWELL, E. D., and C. J. EVERETT, *The Monte Carlo Method For Random Walk Problems*, Pergamon Press, New York, 1959.

DONAHUE, T. M., and G. E. THOMAS, *Planet. Space Sci.*, **10,** 65, (1963).

HEAVENS, O. S., *Optical Properties of Thin Solid Films*, Butterworth's, London, 1955.

HELGE-PETERSON, *Pub. Danske Meteorologiske Institut*, Copenhagen, No. 6, 1928.

HERZBERG, G., *Spectra of Diatomic Molecules*, Van Nostrand, Princeton, 1950.

HIRSCHFELDER, J. O., C. F. CURTISS, R. B. BIRD, *Molecular Theory of Gases and Liquids*, John Wiley, New York, 1954.

KAHN, H., in *Symposium on Monte Carlo Methods*, John Wiley, New York, 1956.

KENNARD, E. H., *Kinetic Theory of Gases*, McGraw-Hill, 1938, p. 286.

MANGE, P., *Ann. Geophysique* **17,** 277, (1961).

MITRA, S. K., *The Upper Atmosphere*, The Asiatic Society, Calcutta, 1952.

NICOLET, M., in *Physics of the Upper Atmosphere* (J. A. Ratcliffe, ed.), Academic Press, New York, 1960.

SCHAAF, S. A., and P. L. CAHMBRE, *Flow of Rarefied Gases*, Princeton U. Press, 1961.

SPITZER, L. JR., in *The Atmosphere of the Earth and Planets*, (G. P. Kuiper, ed.) Univ. of Chicago Press, Chicago, 1952.

UREY, H. C., in *Handbuch der Physik*, Vol. LII, The Solar System, Springer-Verlag, 1959.

ACOUSTIC HEATING OF THE POLAR NIGHT MESOSPHERE

KAICHI MAEDA

Goddard Space Flight Center

As a part of the dynamical heating of the upper atmosphere, which has been emphasized by Hines, the acoustic heating of the upper air is reinvestigated, particularly the warming of the polar night mesosphere.

It is shown that because of the relatively small transmissivity of the thermosphere for long period sound waves, most acoustic heating does not take place above the 200 km level, but rather below 100 km or around the mesopause, and that although the acoustic output power from the polar jet stream is more than 10^5 times larger in winter than in summer, acoustic heating is not sufficient to compensate for the cooling rate around the mesopause in winter, unless the wind velocity of the polar night jet stream continually exceeds 200 m/sec.

INTRODUCTION

An effective heating mechanism for the polar mesosphere, which is warmer in winter than in summer, is Kellogg's chemical heating through the recombination of atomic oxygen.[1–3] Although the velocity of subsidence required for this process is significantly less than that required for adiabatic heating alone, it must still be of the order of 0.1 cm/sec near the 60–90 km level.

On the other hand, theoretical calculations by Haurwitz[4] indicate an upward rather than a downward flow in this region. And although Sawada (1961) has questioned Haurwitz's omission of certain inertia terms in his basic equations, a recent analysis by Kochanski[5] of rocket data relating to mesospheric circulation shows good agreement with Haurwitz's results.

Rocket soundings by Nordberg and Smith[6] have occasionally measured very warm wintertime mesosphere temperatures even in the middle latitudes, where subsidences are unlikely but very strong westerlies in the 40–80 km region appear to be associated.

Hines[7,8] has suggested that the increase of temperature in the polar night mesosphere could be explained by dynamical heating. Gassard[9] and Blamont and Jage[10] have found evidence of upward propagation of tropospheric disturbances into the ionosphere.

The possibility of acoustic heating of the ionosphere was discussed by Daniel,[11] who considered the noise produced by the sea surface as an energy source. However, Eckart[12] showed that pressure waves from this source are completely ineffective in transferring energy up to such high altitudes. This has recently been confirmed by Cook and Young.[13] The amount of acoustic heating of the ionosphere is also estimated by Golitsyn,[14] using tropospheric sources. He arrives at an upward energy flux of the order of 10^{-2} erg/cm^2 sec, which is insufficient to cause the observed heating. Moreover, Golitsyn did not consider the reflection of acoustic energy due to the vertical temperature distribution in the atmosphere.

In this paper, the contribution of aerodynamically produced atmospheric acoustic waves to the warming of the polar night mesosphere is reexamined, with the polar night jet stream in the stratosphere as the source, and with the thermal structure of the polar upper atmosphere, playing a part in propagation and absorption.

It should be noticed that the largest portion of the kinetic energy of the atmosphere is in the

troposphere, where the air density is greatest, and is of the order of 10^3 ergs/cm^3. If this tropospheric kinetic energy is not trapped effectively within the troposphere, the earth's atmosphere at ionospheric heights should reach very high temperatures of the order of 10^5°K. However, Charney and Drazin[15] have shown that the escape of large amounts of planetary wave energy from the tropospheric layer into the upper atmosphere is prevented throughout most of the year by the large westerly zonal wind systems above the tropopause, except for short periods in the spring and autumn.

KINETIC ENERGY DENSITY OF THE POLAR STRATOSPHERE

Owing to the development of the polar night vortex and jet stream, the kinetic energy of the polar stratosphere increases remarkably during the polar nights. As may be the case for the solar corona, some part of this kinetic energy may be transferred upward by means of internal pressure waves.

According to Boville, Wilson and Hare,[16] the kinetic energy density of zonal wind systems between 60°N and 80°N at 25 and 500 mb increases significantly during the wintertime (from December until March). The kinetic energy density corresponding to the winter peak of the basic zonal flow is of the order of 200 and 10^3 ergs/cm^3 at the 25 and 500 mb levels, respectively.

Figure 1 shows monthly averages[17] of kinetic energy densities during the period from July 1957 to July 1958 at 30 mb (solid lines) and at 50 mb (dashed lines) above Ft. Churchill (59°N geographic). This increase of kinetic energy in the upper stratosphere in wintertime is mainly due to the development of the polar night jet stream.[18]

The dissipation of the energy of jet streams, as well as any strong wind system in the atmosphere, is mainly due to friction in the boundary layers, which cause the turbulent motion to decay into small eddies in the surrounding mediums. These eddies decay further into thermal motion and cause local heating. However, some part of the kinetic energy in turbulent motion in jet flow can propagate into the upper atmosphere in the form of pressure waves. Owing to the exponential decrease of air density with height, the amplitudes of those propagating pressure waves grow rapidly at high altitudes and approach shock-wave magnitudes, i.e. dissipation becomes non-linear.

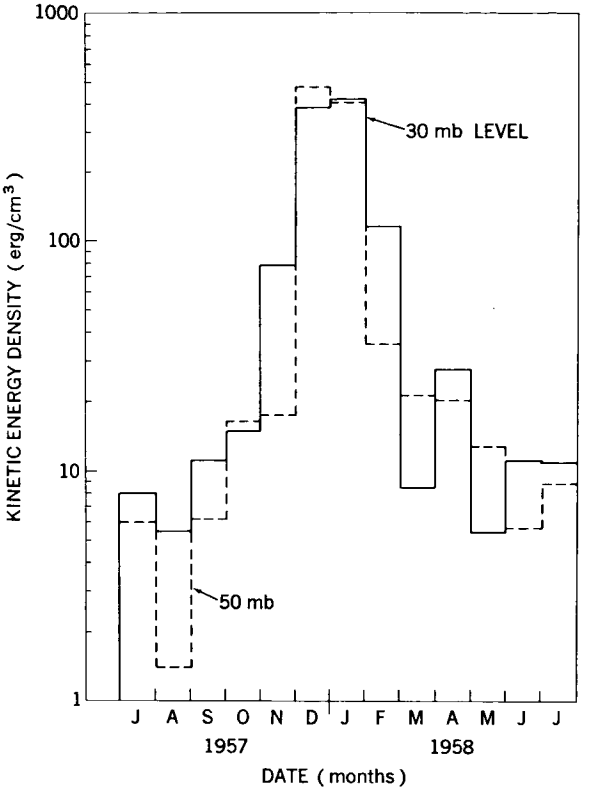


FIGURE 1.—Monthly averages of kinetic energy density above Ft. Churchill (59°N geographic latitude) at 30 mb (solid lines) and at 50 mb (dashed lines) during the period from July 1957 to July 1958.

THE EQUATION OF INTERNAL ATMOSPHERIC WAVES

Fundamental Equations

The equation of internal atmospheric waves is derived from the following three fundamental equations: the equation of motion

$$\frac{D\vec{v}}{Dt} + \frac{1}{\bar{\rho}}\bar{\nabla}\bar{p} + \vec{g} = \vec{f}\,; \qquad (1)$$

the equation of continuity

$$\frac{D}{Dt}\left(\frac{1}{\bar{\rho}}\right) = \frac{1}{\bar{\rho}}\vec{\nabla}\cdot\vec{v}\,; \qquad (2)$$

and the equation of thermodynamics

$$\frac{D\bar{p}}{Dt} = c^2 \frac{D\bar{\rho}}{Dt} . \qquad (3)$$

Since the source of disturbance is extended along a certain latitude, we can treat the problem in two dimensions and specify the coordinates within a meridional plane. Let x be the horizontal (southward) and z the vertical (upward).

In the first approximation, the Coriolis force and external forces, except gravity, are neglected. The equation of motion (equation 1) then becomes

$$\rho_0 \frac{\partial u}{\partial t} = -\frac{\partial p}{\partial x}$$

and

$$\rho_0 \frac{\partial w}{\partial t} = -\frac{\partial p}{\partial z} - g\rho .$$

The equation of continuity (equation 2) for the first order approximation is

$$\frac{\partial \rho}{\partial t} + w \frac{\partial \rho_0}{\partial z} = -\rho_0 \chi , \qquad (4)$$

where $\chi = \chi(x, z, t)$ is the velocity divergence as defined in the previous section. Using equation 4, the equation of thermodynamics (equation 3) can be written as

$$\frac{\partial p}{\partial t} + (c^2 \chi - wg)\rho_0 = 0$$

The waves travelling in the flat atmosphere can be found by assuming that u, w, p and ρ are proportional to $\exp[i(\omega t - kx)]$. Then, the equation for vertical change of $\chi(\omega, z)$ is given by eliminating u, w, p and ρ:

$$\frac{d^2\chi}{dz^2} + \frac{1}{c^2}\left(\frac{dc^2}{dz} - \gamma g\right)\frac{d\chi}{dz} + \left[\frac{\omega^2}{c^2} - k^2 + \frac{g^2 k^2}{\omega^2 c^2}\left(\frac{dc^2}{dz} + (\gamma - 1)g\right)\right]\chi = 0 .$$

Since $c^2 = \gamma g H$, this can be written as

$$\frac{d^2\chi}{dz^2} + \frac{1}{H}(H' - 1)\frac{d\chi}{dz} + \left[\frac{\omega^2}{\gamma g H} - k^2 + \frac{k^2 g^2}{\omega^2 H}\left(H' + 1 - \frac{1}{\gamma}\right)\right]\chi = 0 , \qquad (5)$$

where $H' = dH/dz$.

The Diagnostic Diagram

The wave equation (equation 5) can be written

$$\frac{d^2\chi}{dz^2} - 2N\frac{d\chi}{dz} + M^2\chi = 0 , \qquad (6)$$

where

$$N = \frac{1}{2H}(1 - H') , \qquad (7)$$

$$M^2 = \frac{\omega^2}{c^2} - k^2 + \frac{k^2 g}{\omega^2 H}\left(H' + 1 - \frac{1}{\gamma}\right) \qquad (8)$$

The solution of the differential equation (equation 6) is given approximately by

$$\chi(\omega, z) = e^{Nz}\left(Ae^{-nz} + Be^{+nz}\right) , \qquad (9)$$

where A and B are constant and can be determined by boundary conditions.

It should be noted that the solution (equation 9) of equations 5 or 6 can be classified into cellular

and non-cellular types*, depending upon the relative magnitudes of N and M, i.e.,

(i) cellular solution

$$n = i\eta , \quad \eta^2 = M^2 - N^2 \quad \text{for} \quad M^2 > N^2 , \tag{10}$$

(ii) non-cellular solution

$$n^2 = N^2 - M^2 \quad \text{for} \quad N^2 < M^2 . \tag{11}$$

The cellular waves have propagation vectors in both the horizontal (k) and vertical (η) directions, but non-cellular wave has no propagation vector in the vertical direction and simply propagates in the horizontal direction.

Since M is a function of the horizontal wave number, k, equation 10 gives the relation between two wave numbers in the vertical (η) and the horizontal direction (k): equation 10 is written as

$$\frac{\eta^2 c^2}{\omega^2 - \omega_A^2} + \frac{k^2 c^2 (\omega^2 - \omega_B^2)}{\omega^2 - \omega_A^2} = 1 , \tag{12}$$

where

$$\omega_A^2 = \left(\frac{\gamma g}{2c^2}\right)^2 \left(1 - \frac{1}{\gamma g}\frac{dc^2}{dz}\right)^2 \text{ or } \omega_A = \frac{c}{2H}(1 - H') , \tag{13}$$

$$\omega_B^2 = \frac{g}{c^2}\left(\frac{dc^2}{dz} + (\gamma - 1)g\right) \text{ or } \omega_B = \sqrt{g}\sqrt{\frac{H'}{H} + \frac{1}{H}\left(1 - \frac{1}{\gamma}\right)} ; \tag{14}$$

ω_A and ω_B are called the critical acoustic angular frequency and the Brunt–Väisälä angular frequency. These are both functions of the atmospheric temperature distribution in the vertical direction and are related to the stability of layers in the atmosphere.

Since $\omega_A > \omega_B$ we can see from equation 12 that, if $\omega > \omega_A$, the propagation surface is an ellipsoid while if $\omega < \omega_B$, it is an hyperboloid. In both cases, the rotation axis (z-axis) is vertical (figure 2). The waves corresponding to $\omega > \omega_A$ are called sonic or acoustic waves and the ones for $\omega < \omega_B$ are called thermobaric waves[19] or internal gravity waves[20] respectively.

Since η^2 becomes negative for $\omega_B < \omega < \omega_A$, no cellular wave exists for this frequency range. This situation can be seen from figure 3, in which are plotted as a function of k the two curves of ω, corresponding to $\eta^2 - n^2 - 0$. Similar curves are shown in figure 4, where the wave period (τ in sec) is plotted against the horizontal wave length instead of ω and k, respectively. τ_A and τ_B correspond to ω_A and ω_B, respectively.

Figures 3 and 4 are called diagnostic diagrams of atmospheric waves,[21] in which a domain indicated by $n^2 > 0$ corresponds to the non-cellular

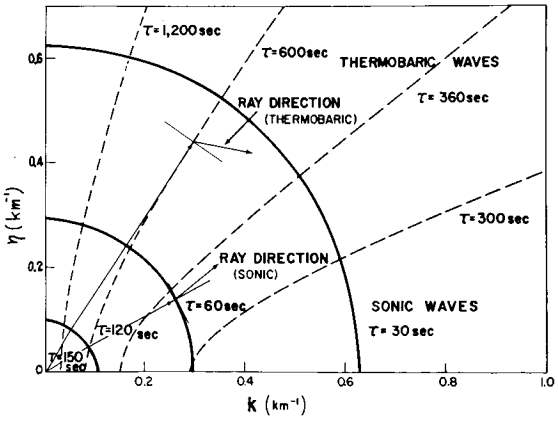


FIGURE 2.—Propagation surfaces (lines) of internal atmospheric waves in the isothermal atmosphere with scale height 8 km ($T = 0°K$), given by equation 12; the solid and dashed lines correspond to sonic waves (acoustic waves) and thermobaric waves (internal gravity waves), respectively.

* The nomenclature of cellular and non-cellular is taken from Martyn's paper (reference 21) and this corresponds to internal and external waves discussed by Charney and Drazin (reference 20).

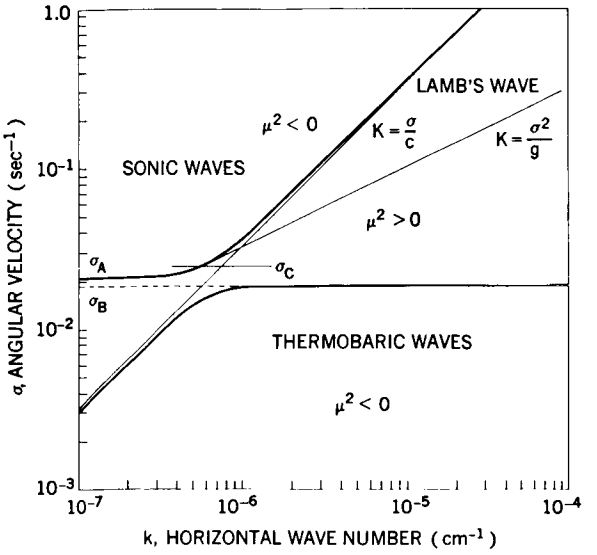


FIGURE 3.—Diagnostic diagram of internal atmospheric waves in the isothermal atmosphere with scale height 8 km ($T = 0°K$).

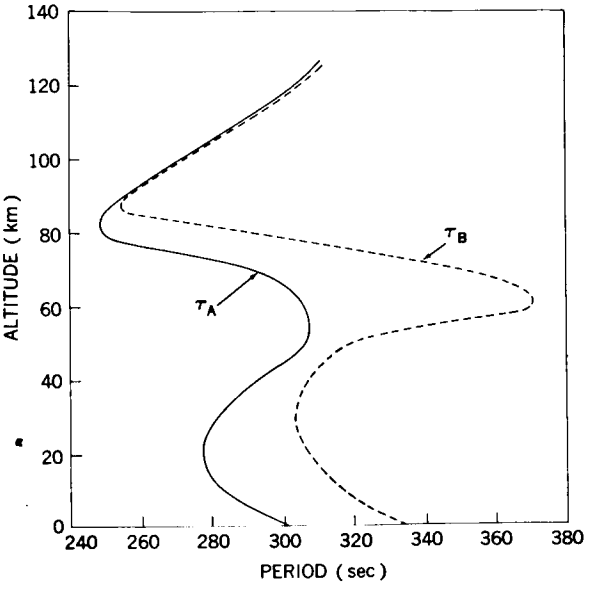


FIGURE 5.—Critical acoustic period, τ_A, and Brunt-Väisälä period, τ_B, as a function of altitude for the summer polar atmosphere.

mode while two domains $n^2 < 0$ correspond to the cellular modes.

APPROXIMATE TRANSMISSIVITY OF THE POLAR ATMOSPHERE

Since the non-cellular (or external) waves, the period of which is limited between τ_A and τ_B, have no vertical propagation vector, these waves neither contribute to the vertical energy transfer nor exist locally in free atmosphere. Among two kinds of cellular (or internal) waves, on the other hand, the thermobaric (or internal gravity) mode can propagate with larger periods than the

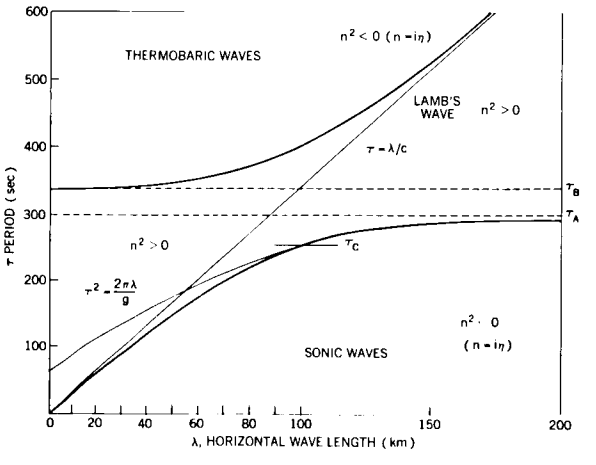


FIGURE 4.—Diagnostic diagram of internal atmospheric waves in the isothermal atmosphere with scale height 8 km ($T = 0°K$).

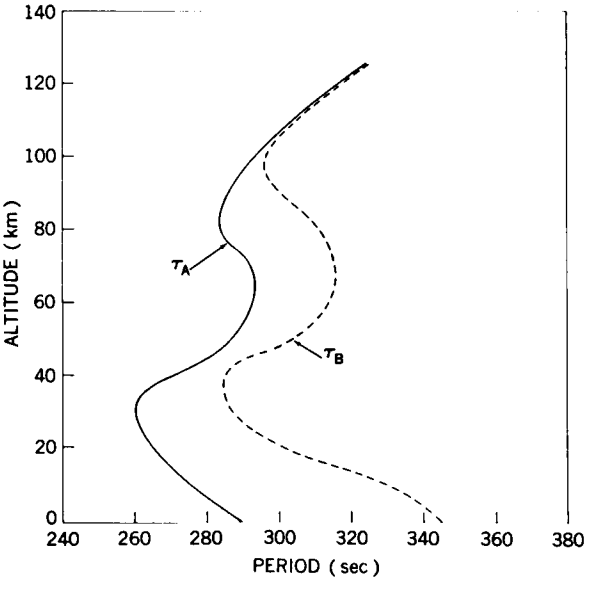


FIGURE 6.—Critical acoustic period, τ_A, and Brunt-Väisälä period, τ_B, as a function of altitude for the winter polar atmosphere.

Brunt period τ_B, while the sonic (or acoustic) modes can exist only below a period τ_A.

As can be seen from equations 13 and 14, these two critical periods are a function of temperature and vertical temperature gradients in the atmospheric temperature). Therefore, these values differ not only with height but also with season. In figures 5 and 6, τ_A and τ_B are plotted against altitude for polar atmospheres above Fort Churchill in summer and in winter, respectively.

The atmospheric models of the two seasons are essentially the same as the ones reported by Stroud et al. (reference 22; see also figure 13, reference 23 and figure 1, reference 2). In a domain between two critical periods, i.e., τ_A (solid curves) and τ_B (dashed curves), only the noncellular mode can exist. In other words, this is a forbidden region for both acoustic and thermobaric modes of internal atmospheric waves. Therefore, we can see that no thermobaric waves with periods shorter than 370 sec can propagate across the 50-km level and that no sonic waves with periods longer than 250 sec can propagate through the layer around 80 km in summertime. But in winter both waves are able to propagate through these layers and the critical period for each approaches 320 sec for thermobaric waves and 290 sec for sonic waves, respectively.

To see the transparency (or transmissivity) of the atmosphere for internal atmospheric waves more quantitatively, the wave equation (equation 5 or 6) should be transformed into a more simple form. By substituting

$$y = \int_{z_0}^{z} \frac{dz}{H}$$

and

$$\chi(z) = e^{y/2}\phi(y) ,$$

the term $d\chi/dz$ in the wave equation can be eliminated; then equation 6 can be written as

$$\frac{d^2\phi(y)}{dy^2} = -K^2\phi(y) \qquad (15)$$

where

$$K^2 = \eta^2 \cdot H^2 .$$

As can be seen from equations 7, 8, and 10, η^2 is a function of σ, k, H, and $H'(=dH/dz)$.

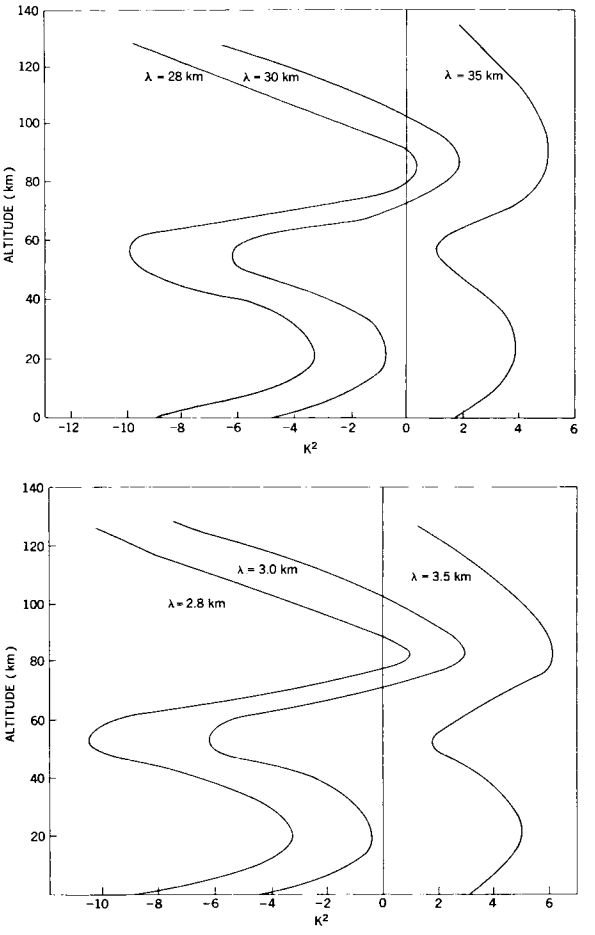


FIGURE 7.—Square of the index of refraction, k, for atmospheric acoustic waves with period τ in the summertime polar atmosphere as a function of height, where the parameter λ is the horizontal wave length in km.

The one-dimensional wave equation (equation 15) indicates that, if $K^2>0$, the solution $\phi(y)$ has a wave nature and the corresponding waves can propagate along the y-axis (z-axis vertical direction) within the domain, where this condition is satisfied. On the other hand, if $K^2\leq 0$ (K is imaginary) i.e., no internal wave exists and vertical propagation of waves is inhibited by the barrier, where K^2 becomes negative. Thus, the propagation of internal waves is limited within regions where K is real.

For the acoustic mode of atmospheric waves with periods 100 sec and 10 sec, K^2 is plotted against z in figures 7a and 7b for summertime and in figures 8a and 8b for wintertime, respectively.

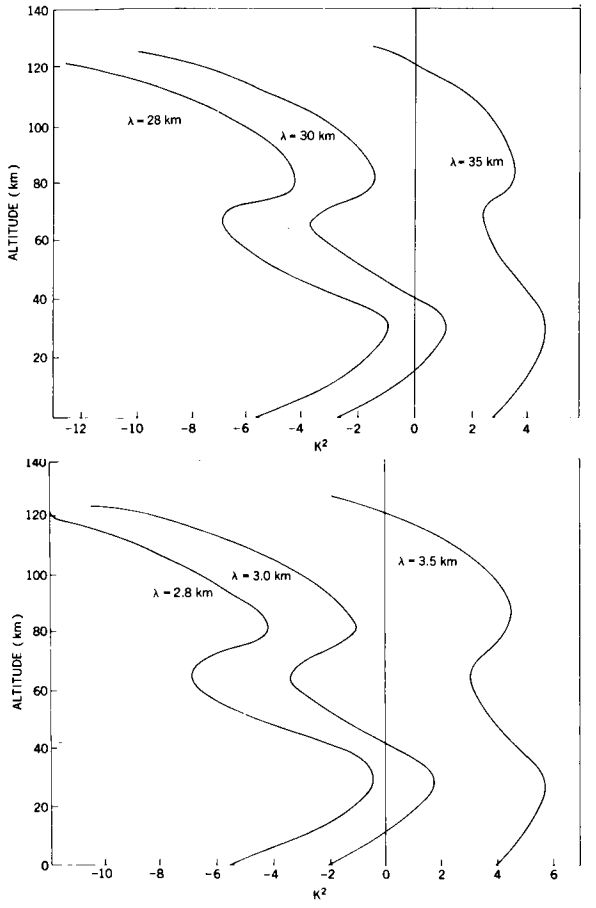


FIGURE 8.—Curves similar to figure 7 for wintertime polar atmosphere.

In these figures, the horizontal wavelength is taken as a parameter.

Since $\lambda = 2\pi/k$, where k is a wave number in the horizontal direction, there is a shortest (critical) horizontal wavelength λ_c for a given period of the waves, corresponding to a largest horizontal wave number k_c. From figures 7 and 8, we can see that the waves propagating close to the vertical axis can propagate upwards as far as $\tau < \tau_A$, while the waves propagating in an oblique direction meet a barrier where K^2 becomes negative even if the condition $\tau < \tau_A$ is satisfied. If there are more than two barriers, those obliquely propagating waves are trapped between these barriers, forming the "duct" in the horizontal direction. It should be noted from figure 2 that $k = 0$ (i.e. $\lambda = \infty$) represents the wave propagation in the vertical direction.

Using the diagram shown in figure 2 for the conditions at each level, the critical angle θ_c can be drawn as a function of altitude for a given period, τ, of the waves, provided that $\tau < \tau_A$. The angle θ_c is defined such that the waves propagating obliquely with zenith angle larger than θ_c cannot enter into the above layer, simply because K^2 becomes negative beyond this angle. In other words, θ_c can be regarded as the aperture of the allowed cone around the vertical axis at each level in the atmosphere. Therefore, the relative transmissivity, $f_t(\tau, z)$, of each layer for vertical propagation of the atmospheric acoustic wave is approximately given by $(\theta_c/\pi/2)^2$, provided that θ_c is expressed in radians. In figure 9, $f_t(\tau, z)$ is shown for the wave with a 100 sec period normalized to the condition at 25 km. Similar curves are obtainable for any waves with $\tau < \tau_A$. From figure 9, we can see that the atmosphere above

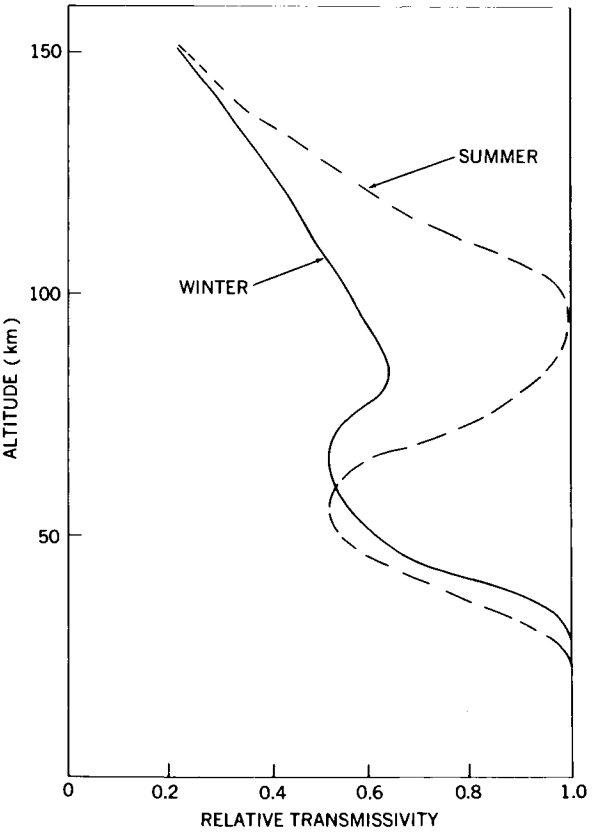


FIGURE 9.—Relative transmissivity of the polar atmosphere, $f_t(\tau, z)$ for atmospheric acoustic waves with a period $\tau = 100$ sec. The solid and dashed lines correspond to wintertime and summertime, respectively.

mesopause, which is nearly 80 km in polar regions, is always opaque for atmospheric acoustic waves.

It should be noted that the wind system in the higher layers does not affect $f_t(\tau, z)$ very much, because an oblique axis tilted windward direction can be regarded as the direction of maximum propagation.

UPWARD ENERGY TRANSFER BY PRESSURE WAVES

Acoustic Waves Generated Aerodynamically in the Atmosphere

Sound waves generated aerodynamically are distinct from sound produced by the vibration of solids. A strong jet flow in the free atmosphere produces sound waves due to turbulent motions in the boundary layers between the core of the jet and the surrounding air. Lighthill[24,25] investigated this problem and derived a theoretical expression of the power of sound waves as a function of the jet stream velocity. His theory was further extended by Proudman[26] to the generation of sonic waves by isotropic turbulence.

Although the acoustic heating of the solar corona is discussed by Schatzman[27] similar calculations cannot be applied to the earth's atmosphere, because the situation there is entirely different from that of solar atmosphere. In the earth's upper atmosphere, the effect of magnetic fields on the generation of sound noise by isotropic turbulence[28] is important as well as their effect on the dissipation processes as in the case of solar corona.

Since the aerodynamical production of acoustic waves has been developed in detail by Lighthill and Proudman, only their results will be given, and, in turn, applied to the present problem in the earth's upper atmosphere.

The acoustic power output per unit mass of air P, in ergs/gm-sec, is given by

$$P = \alpha \epsilon M^5 \quad \text{for} \quad M \ll 1 \qquad (16)$$

where α is a nondimensional constant, M is Mach number ϵ is the mean rate of energy dissipation per unit mass.

According to Batchelor (reference 29, p. 103), ϵ is given by

$$\epsilon \left(\approx \frac{3}{2} \frac{du^2}{dt} \right) = A \frac{\bar{u}^3}{l} ,$$

where A is a nondimensional constant, which is of the order of unity—about 1.8, l a characteristic length which can be taken as the reciprocal of the wave number at which the maximum of the energy spectrum occurs, and u the root mean square velocity of the turbulence.

If the "double correlation function" given by Heisenberg[30] is correct, the constant α in equation 16 has the value 38. Therefore equation 16 is

$$P = 70 \frac{\bar{u}^3}{l} \cdot (M)^5$$

where $M = u/c$, u and c are the mean velocity of jet flow and the sound velocity respectively.

In the upper atmosphere, where the Reynolds number is very large, the spectrum of turbulence consists generally of the following ranges:

1. The energy-producing range, which contains large eddies.
2. The inertial subrange in which energy is neither created nor destroyed.
3. The energy-dissipating range, which contains small eddies.

As pointed out by Proudman,[26] the turbulence which makes an appreciable contribution to the generation of sonic waves at large Reynolds numbers belongs to the eddies that do not dissipate energy. In other words, the main source of aerodynamically produced sonic waves is turbulent motion around the jet flow; this turbulence belongs to (1) energy producing range, and to (2) inertial sub-range. Therefore, following Batchelor (reference 29, Chap. 7), we can assume

$$10^3 \lesssim l \lesssim 10^4$$

and

$$\frac{1}{10} \lesssim \frac{\bar{u}}{u} \lesssim \frac{1}{3} .$$

Distribution of turbulence along the jet stream can be assumed uniform, then the acoustic output per unit area along the jet (in erg/cm^2-sec) will be

$$7 \cdot 10^{-6} u^3 M^5 \rho_0 \lesssim I(u) \lesssim 3 \cdot 10^{-3} u^3 M^5 \rho_0 , \qquad (17)$$

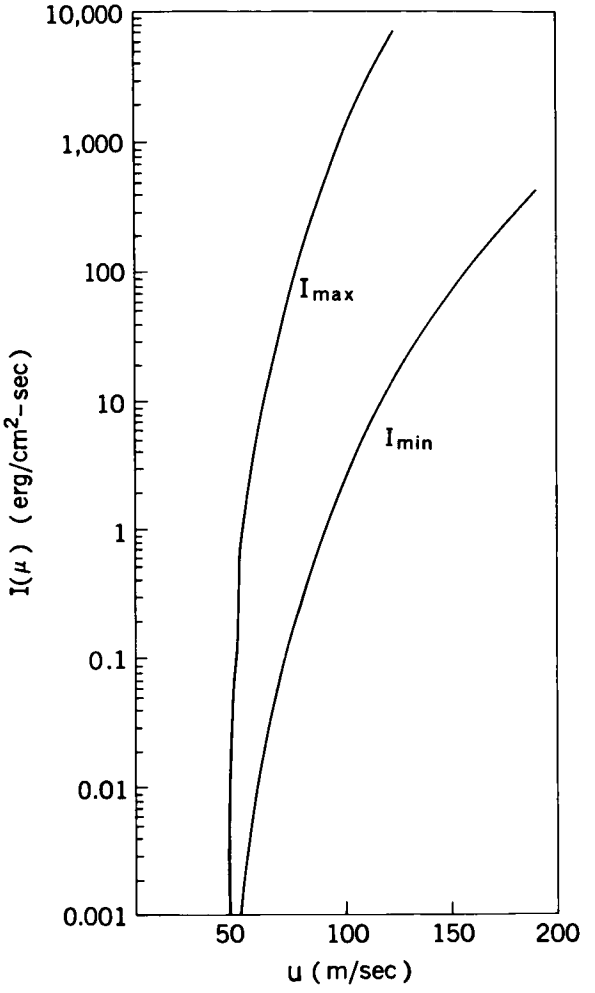


FIGURE 10.—Acoustic output power of the polar stratosphere, $1(u)$, plotted against the mean wind velocity of the jet stream, u.

where ρ_0 is the mean density of air around the source. In equation 17, u is in cm/sec. The acoustic output $I(u)$ is shown in figure 10, assuming the height of the source to be 25 km.

From this figure we can see that the production of sonic waves decreases drastically below a wind velocity of the order of 50 m/sec. Since the derivation of equation 17 is based on the assumption that Mach number is small,[25] the output power beyond $u=150$ m/sec might be incorrect.

Attenuation of Atmospheric Waves

Any atmospheric wave is attenuated in air by the following effects:

1. friction due to the viscosity of air.
2. thermal conduction of air.
3. radiation through the atmosphere.
4. dispersion due to the inhomogeneity of air in the atmosphere.

As was shown by Rayleigh,[31] effects of radiation and dispersion are generally small for the acoustic waves as compared with the effects of friction and thermal conduction.

Consideration of these effects leads to the so-called classical Stokes–Kirchhoff formula for the attenuation coefficient $\alpha(\tau)$,[31] which is given by

$$\alpha(\tau) \approx \frac{4\pi^2}{\tau^2 c^3}\left[\frac{\gamma-1}{\gamma}a^2+\frac{4}{3}\nu\right], \quad (18)$$

where α^2 and ν are the coefficients of thermal conductivity and kinematic viscosity of air:

$$\alpha^2 = \frac{\kappa}{c_p \rho}$$

and

$$\nu = \frac{\mu}{\rho}$$

The heat conductivity of air, κ, is the order of 6×10^{-6} cal/cm-sec°K at NTP; and the specific

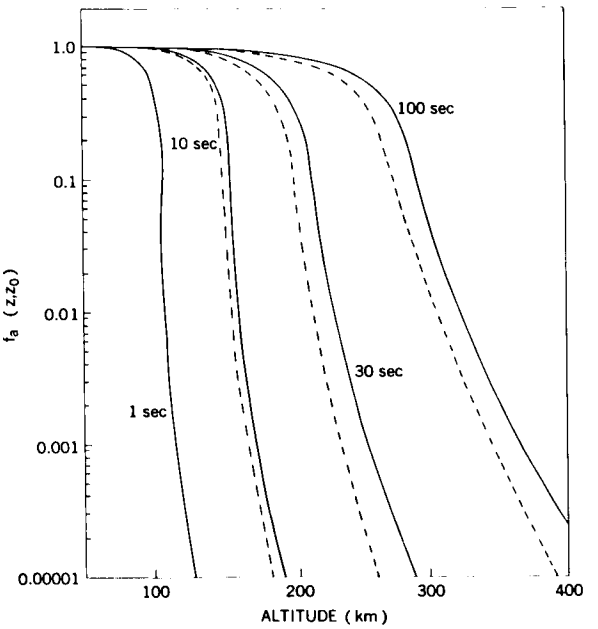


FIGURE 11.—Attenuation factor, $f_a(\tau, z)$, of atmospheric acoustic wave vs. altitude, z for $\tau=1$, 10, 30, and 100 sec. The solid lines are calculated by the classical formula and the dashed lines are given by Golitsyn, with the Joule loss in the ionosphere taken into account.

heat C_p is nearly 0.27 cal/gm°K below the mesopause and increases with height to 0.3 cal/gm°K above 350 km approximately.[32]

The viscosity of air, μ, is of the order of 1.7×10^{-4} poise at NTP. Using these numerical values, the attenuation coefficient (equation 18) becomes

$$\alpha(\tau) \approx \frac{2.4 \times 10^{-16}}{\rho\tau^2}$$

As was shown by Golitsyn,[33] the long period pressure wave undergoes an additional dissipation owing to the Joule-loss of the current induced by fluctuations of the conducting medium of ionosphere in the earth's magnetic field.

In figure 11, the attenuation factor

$$f_a(\tau, z) = \exp\left[-\int_0^z \alpha(\tau, z')\,dz'\right] \quad (19)$$

is shown for four different periods of waves (1, 10, 30 and 100 sec). The dashed lines in figure 11 are taken from Golitsyn's results, which took account of the Joule loss.[33]

An estimate of the rate of energy dissipation per unit time per unit thickness per unit cross section in the upper atmosphere, dE_t/dz in erg/cm³-sec, is made by the following approximation:

$$\frac{dE_t}{dz} \approx \int_{\omega_A(z)}^{\infty} i(\omega)\, f_t(\omega, z)\, \frac{d}{dz} f_a(\omega, z)\, d\omega\,, \quad (20)$$

where $f_a(\omega, z)$, the attenuation factor, is given by equation 19 and shown in figure 11; $f_t(\omega, z)$ is the relative transmissivity of atmosphere for the wave with angular frequency ω at the altitude z, which is shown, for example, in figure 9 for $\tau = (2\pi/\omega) = 100$ sec; and $i(\omega)$ is the differential frequency spectrum of the wave, which will be assumed to be

$$i(\omega) = i_0 \omega^{-m} \qquad m > 1\,. \quad (21)$$

From figures 5 and 6, the lower limit ω_A of the integral in equation 20 can be taken as follows:

(i) summer

$\omega_A = \omega_A(z)$ for $0 \le z < 20$ km,
0.023 for $20 \le z < 70$ km,
$\omega_A(z)$ for $70 \le z < 85$ km,
0.025 for $z \ge 85$ km;

(ii) winter

$\omega_A = \omega_A(z)$ for $0 \le z < 30$ km,
0.024 for $z \ge 30$ km.

From this figure, we can see that the change of ω_A with height is small. Thus, ω_A is taken as a constant value (0.024) for both seasons.

Strictly speaking, the right hand side of equation 20 should be multiplied by a geometrical factor, which is inversely proportional to the distance from the source. The width of the source —the polar night jet stream—is more than several hundred kilometers[18] and the boundary is widely diffused horizontally. In this respect, the geometric attenuation factor is not important and can be neglected in the present approximation.

Since the waves are propagating upwards, $f_t(\omega, z)$ in the integrand of equation 20 may be taken as a constant value above the minimum transmissivity level. However, the formation of wave-ducts is highly probable above this or between those levels of minimum transmissivity, for example, between 30 and 70 km in summer (see figure 5). As can be seen from the result, the choice of $f_t(\omega, z)$ above the minimum transmissivity level (if it is constant or variable as shown in figure 9, for example), does not affect the present calculation very much.

A constant i_0 in equation 21 is given by

$$I_0 = \int_{\omega_A(z_0)}^{\infty} i_0 \omega^{-m}\, d\sigma = i_0 \frac{\omega_A^{m-1}}{m-1} \quad (22)$$

where $I_0 = I_0(u)$ is the output power of the acoustic wave at the level of the source, z_0, given by equation 17 and shown in figure 10.

In figure 12, the relative dissipation rate of energy $(dE_t/dz)/I_0$ is plotted against altitude z

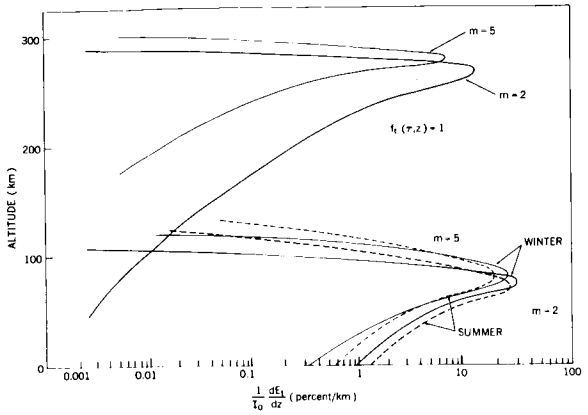


FIGURE 12.—Relative dissipation rate of energy flux transferred by atmospheric acoustic waves, where $I_0 = I_0\ (u)$ is the output at the source as shown in figure 10. Solid and dashed lines correspond to wintertime and summertime polar atmospheres, respectively; heavy and light lines stand for the hard spectrum ($m=2$) and soft spectrum ($m=5$) of the waves, respectively.

for two cases of the spectrum, $m=2$ (heavy lines) and $m=5$ (light lines). The solid and dashed lines correspond to the wintertime and the summertime polar atmosphere, respectively. The two curves in the upper portion of figure 12 are calculated by neglecting the transmissivity of the atmosphere, i.e. $f_t(\omega, z)$ is assumed to be unity. The latter corresponds to the estimation made by Golitsyn.[14]

If all those dissipated energies of acoustic pressure waves are assumed to be converted into thermal energy of the air, the rate of heating at each level, $dT(z)/dt$ in °C/sec, is given by

$$\frac{dT}{dt} \approx \frac{1}{C_p \rho(z)} \cdot \frac{1}{J_0} \frac{dE_t}{dz} , \qquad (23)$$

where J_0 is Joule's mechanical equivalent of heat, 4.2×10^7 ergs/cal; C_p is the specific heat of air at constant pressure, ~0.3 cal/gm °C; and $\rho(z)$ is the mean atmospheric density at the altitude z.

The results are shown in figure 13, which corresponds to figure 12 for a unit output power source. Since the latter is shown in figure 10, the following points may be observed:

If the wind velocity around 25 km level is of the order of 30 m/sec, as in the usual summertime polar stratosphere, we can see that the acoustic power at the source is less than the order of 10^{-5} erg/cm²-sec (from figure 10). Then rate of maximum heating, dT/dz is, from figure 13, of the order of $10^{-8}\times10^{-5}$ °C/sec or 10^{-8} °C/day at the maximum, which height is around 75 km if the wave spectrum is hard ($m=2$), or 100 km if the spectrum is soft ($m=5$), respectively.

On the other hand, if the wind velocity around 25 km exceeds 100 m/sec, which is steadily observed in the polar night jet stream,[17,34] the output power I_0 is somewhere between 10^3 and 1 erg/cm² sec (figure 10). The corresponding maximum warming rate is, then $2\times(10^{-5}$ to $10^{-8})$ °C/sec or $2\times(1$ to $10^{-3})$ °C/day at the altitude of 75 km for $m=2$, or around 100 km for $m=5$.

If the transmissivity of the atmosphere is disregarded, as in Golitsyn's estimation, then from the two upper curves shown in figure 13, the rate of maximum heating is of the order of 10^{-10} °C/sec (10^{-5} °C/day) in summer, and 10^{-2} to 10^{-5} (10^3 to 1°C/day) in winter. The height of maximum heating is around 260 to 300 km which corresponds to $m=2$ to 5.

CONCLUSION

Although the estimated acoustic energy flux available for upper air heating has a wide range as shown in figure 10—due to ambiguities of

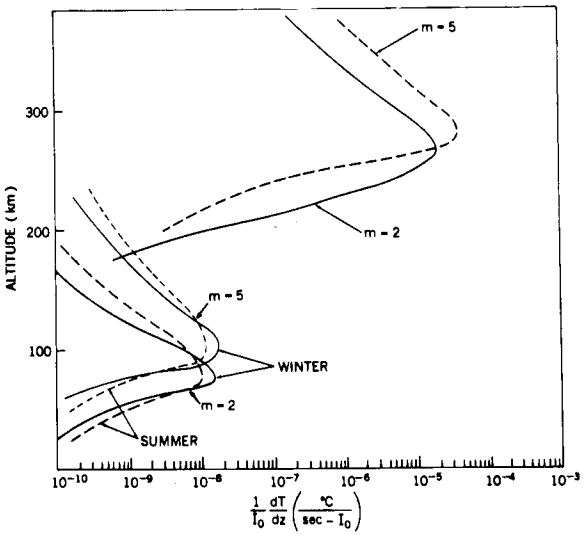


FIGURE 13.—The rate of atmospheric heating in each layer due to the dissipation of acoustic energy flux I_0 (erg/cm² sec) vs. altitude z. The lines are defined as in figure 12.

several parameters such as the mean velocity of turbulent motion around the jet stream, $\bar{u}$; the effective size of eddy in this turbulence, l; and the frequency spectrum of the turbulence, the following remarks can be made from the present calculation.

Because of the temperature gradient, the atmosphere above the 100 km level is not transparent for atmospheric acoustic waves—especially for long waves. Therefore, acoustic heating is most effective for warming around the 100 km level rather than the 200 to 300 km levels (figure 13).

On the other hand, if the transmissivity of sound waves due to thermal gradient in the atmosphere is disregarded, most heating takes place around the 300 km level and its rate exceeds 10°C/day in winter.

Although acoustic power from the polar night jet stream is more than two orders of magnitude larger than that of the tropospheric disturbance, which is of the order of 10^{-2} ergs/cm^2 sec,[14] acoustic heating around the mesopause is not sufficient to compensate for the cooling rate (about 10°C/day) of these levels in winter.

In this respect, another mechanism of dynamical heating such as energy transfer by means of internal gravity waves[8] might be effective for the warming of the mesosphere observed by Nordberg and Smith.[6]

If the wind velocity of the jet stream exceeds 200 m/sec for several days, the layer around the mesopause can be warmed significantly even by acoustic waves.

REFERENCES

1. Kellogg, W. W., "Chemical Heating above the Polar Mesopause in Winter," *J. Meteorol.* **18**(3): 373–381, June 1961.
2. Young, C., and Epstein, E. D., "Atomic Oxygen in the Polar Winter Mesosphere," *J. Atmos. Sci.* **19**(6): 435–443, November 1962.
3. Maeda, K., "Auroral Dissociation of Molecular Oxygen in the Polar Mesosphere," *J. Geophys. Res.* **68**(1): 185–197, January 1, 1963.
4. Haurwitz, B., "Frictional Effects and the Meridional Circulation in the Mesosphere," *J. Geophys. Res.* **66**(8): 2381–2391, August 1961.
5. Kochanski, A., "Circulation and Temperatures at 70- to 100-Kilometer Height," *J. Geophys. Res.* **68**(1): 213–226, January 1, 1963.
6. Nordberg, W., and Smith, W., "Preliminary Measurements of Temperatures and Winds above 50 km over Wallops Island, Virginia," NASA Technical Note D-1694, March 1963.
7. Hines, C. O., "Comments" Annual AGU Meeting, April 1962.
8. Hines, C. O., "The Upper Atmosphere in Motion," *Quart. J. Roy. Meteorol. Soc.* **89:** 1–42, January, 1963.
9. Gossard, E. E., "Vertical Flux of Energy into the Lower Ionosphere from Internal Gravity Waves Troposphere," *J. Geophys. Res.* **67**(2): 745–757, February 1962.
10. Blamont, J. E., and De Jage, C., "Upper Atmospheric Turbulence near the 100 km Level," *Annales de Geophys.* **17**(1): 134–144, 1961.
11. Daniels, F. B., "Acoustical Energy Generated by the Ocean Waves," *J. Acoust. Soc. Am.* **24:** 38, 1952.
12. Eckart, C., "The Theory of Noise in Continuous Media," *J. Acoust. Soc. Am.* **25**: 195–199, March 1953.
13. Cook, R. K., and Young, J. M., "Microbaroms and Sound Radiated by Ocean Waves," Annual AGU Meeting, Washington D. C., April 19, 1963.
14. Golitsyn, G. S., "Possibility of the Heating of the Upper Atmosphere by Long Wave Acoustic Radiation," *Akad. Nauk USSR, Izvest. Ser. Geofiz.* **7:** 1092–1093, July 1961; Translated in *Bull. Acad. Sci. USSR, Geophys. Ser.* **7:** 720–721, 1961.
15. Charney, J. G., and Drazin, P. G., "Propagation of Planetary-Scale Disturbances from the Lower into the Upper Atmosphere," *J. Geophys. Res.* **66**(1): 83–108, January 1961.
16. Boville, B. W., Wilson, C. V., and Hare, F. K., "Baroclinic Waves of the Polar Night Vortex," *J. Meteorology* **18**(5): 567–580, October 1961.
17. U. S. Weather Bureau, "Daily 100-mb and 50-mb Three Times Monthly 30-mb Synoptic Weather Maps of the IGY period," Washington: U. S. Department of Commerce, 1961.
18. Krishnamjrti, T. N., "A Vertical Cross Section Through the 'Polar-Night' Jet Stream," *J. Geophys. Res.* **64**(11): 1835–1844, November 1959.
19. Eckart, C., "Hydrodynamics of Oceans and Atmospheres," New York: Pergamon Press, 1960.
20. Hines, C. O., "Internal Atmospheric Gravity Waves at Ionospheric Heights," *Canad. J. Phys.* **38**(11): 1441–1481, 1960.
21. Martyn, D. F., "Cellular Atmospheric Waves in the Ionosphere and Troposphere," *Proc. Roy. Soc.* **A201:** 216–233, November 1960.
22. Stroud, W. G., Nordberg, W., and Bandeen, W. R., "Rocket-Grenade Measurements of Temperatures and Winds in the Mesosphere over Churchill, Canada," *J. Geophys. Res.* **65**(8): 2307–2323, August 1960.
23. Maeda, K., "On the Heating of the Polar Upper Atmosphere," NASA Technical Report R-141, 1962.
24. Lighthill, M. J., "On Sound Generated Aerodynamically. I. General Theory," *Proc. Roy. Soc.* **A211:** 565–587, March 20, 1952.

25. Lighthill, M. J., "On Sound Generated Aerodynamically. II. Turbulence as a Source of Sound," *Proc. Roy. Soc.* **A222:** 1–32, February 23, 1954.
26. Proudman, I., "The Generation of Noise by Isotropic Turbulence," *Proc. Roy. Soc.* **A214:** 119–132, August 7, 1952.
27. Schatzman, E., "The Propagation of a Shock-Wave in an Atmosphere of Varying Density," *Nuovo Cimento* 22(Serielo): 209–237, 1961.
28. Kulsrud, R. M., "Effect of Magnetic Fields on Generation of Noise by Isotropic Turbulence," *Astrophys. J.* **120**(2): 461–480, March, 1955.
29. Batchelor, G. K., and Bondi, H., "The Theory of Homogeneous Turbulence," Harvard: Cambridge University Press, 1956.
30. Heisenberg, W., "On the Theory of Statistical and Isotropic Turbulence," *Proc. Roy. Soc.* **A195**(1042): 402–406, December 22, 1948.
31. Rayleigh, J. W. S., "The Theory of Sound," 2nd ed., New York: Dover Publications, 1945.
32. Ishikawa, G., "Solar Corpuscular Radiation as a Heat Source of the Upper Atmosphere," *Papers Meteorol. Geophys.* **19**(2): 93–123, December 1959.
33. Golitsyn, G. S., "Sound Absorption in the Atmosphere and Ionosphere," *Akad. Nauk USSR, Izvest. Ser. Geofiz.* **6:** 942–946, June 1961; Translated in *Bull. Acad. Sci. USSR, Geophys. Ser.* **6:** 618–621, 1961.
34. Riehl, H., "Jet Streams of the Atmosphere," Dept. of Atmospheric Science, Colorado State University, Technical Rept. No. 32, May 1962.

Appendix A

NOTATION

$g(0_1-g)$	Acceleration of gravity, $g=98$ cm/sec^2.
γ	Ratio of specific heats of air, $C_p/C_v=1.4$, where C_p and C_v are the specific heat of air at constant pressure and that at constant volume, respectively.
R	Gas constant of air, 2.87×10^6 ergs/mol °C.
J_0	Joule's mechanical equivalent of heat, 4.185×10^7 ergs/cal.
$\vec{f}$	Resultant of all external forces (dyne/gm).
$\vec{v}(u, w)$	Velocity vector, where u is horizontal (southward) and w is vertical (upward) component of air motion in cm/sec.
$\chi(x, z)$	The divergence of velocity per sec, i.e., $\chi = \frac{\gamma u}{\gamma x} + \frac{\gamma \omega}{\gamma z}$, where x and z are horizontal (southward) and vertical (upward) coordinate, respectively.
P, ρ, T	Small departure from static value of pressure, density and temperature, function of x and z.
P_0, ρ_0, T_0	Static pressure, density and absolute temperature, which are the functions of z only in dyne/cm^2, gm/cm^3, and °K, respectively.
$\bar{P}, \bar{\rho}, \bar{T}$	Total pressure, density and absolute temperature, $\bar{P}=P_0+P$, $\bar{\rho}=\rho_0+\rho$ and $\bar{T}=T_0+T$.
κ	Heat conductivity of air (cal/cm sec°K).
λ	Horizontal wave length of pressure wave (km).
τ	Period of pressure wave (sec).
k	Horizontal wave number, corresponding to λ, $k=2\pi/2$ per km.
ω	Angular frequency of pressure wave, corresponding to τ, $2\pi/2$ per sec.
c	Velocity of sound wave in the atmosphere (cm/sec).
H	Scale height of isothermal atmosphere, $H=RT_0/g$.
D/Dt	The Eulerian derivative, $\gamma/\gamma t+\vec{v}\cdot\vec{\nabla}$.
μ	Viscosity of air, in dyne-sec/cm^2 i.e. poise (gm/cm-sec).
ν	Kinematic viscosity of air, $\nu=\mu/\rho_0$ (cm^2/sec).
a^2	Thermal conductivity of air, $a^2=\kappa/C_p\rho$ (cm^2/sec).

AURORAL DISSOCIATION OF MOLECULAR OXYGEN IN THE POLAR MESOSPHERE

KAICHI MAEDA

Goddard Space Flight Center

As had been shown by Kellogg, one of the most effective mechanisms for heating of the polar mesosphere during polar night is the exothermic association of atomic oxygen intensified by the large-scale subsidence of the upper polar atmosphere in winter. Since photodissociation is no longer predominant in the polar regions in winter, the contribution of auroral particles to the dissociation of molecular oxygen is estimated, assuming the atmospheric composition given by Johnson. It is shown that the maximum rate coefficient is of the order of 10^{-8} $(cm^3\ sec)^{-1}$ for protons and electrons of energy spectrum $E^{-n}dE$, where $n \geq 3$, and that the height of maximum dissociation is around 90 km. Consequently, the rate of corpuscular dissociation of oxygen molecules in the upper mesosphere during strong auroral activities is comparable with that of photodissociations, which is mostly due to the Schumann-Runge continuum of solar ultraviolet radiations.

INTRODUCTION

The polar night mesosphere is warmer than any other season and any other latitude even though the heating due to solar radiation is smaller [Sheppard, 1959; Stroud et al., 1960; Kellogg, 1961; Nordberg and Stroud, 1961]. Net cooling around the mesopause, which is mostly due to thermal emission by carbon dioxide, is of the order of 10°C per day, i.e., of the order of 10^3 ergs/g-air sec. To compensate for this cooling by adiabatic heating of subsiding air, quite strong, steady, downward flow—at least 1 km per day, or roughly 2 cm per second—would be necessary.

A better explanation, however, has been given by Kellogg [1961], who showed that the energy release from the recombination of atomic oxygen is one of the most effective heat sources in the mesosphere. He called this process chemical heating. By this process, the subsiding velocity required for the compensation of radiative cooling mentioned above can be reduced to one tenth or less.

The main purpose of this paper is to show the contribution of auroral particles to the heating of the polar mesosphere, a contribution that was ruled out by Kellogg [1961]. It can be shown that direct corpuscular heating of the upper atmosphere by auroral particles is only significant in high altitudes above 100 km, as shown by the estimations given by Krassovsky [1959], Jastrow [1959], and Ishikawa [1959]. Consequently, the effect of direct corpuscular heating on the polar mesosphere is practically negligible.

On the other hand, the dissociation of oxygen molecules in the upper atmosphere, normally caused primarily by solar ultraviolet radiations, can also be caused by auroral particles. Since the photodissociation of polar upper air is no longer predominant during the polar night, the contribution by auroral particles to the corpuscular dissociation of polar upper air cannot be disregarded for mesospheric warming in winter. Part of the energy brought into the polar upper atmosphere by auroral particles can be transferred into the lower atmosphere in the following manner.

As a result of bombardments by auroral particles, molecular oxygen in high altitudes is endothermically dissociated into atomic oxygen. (Nitrogen molecules also are dissociated by electrons and photons. However, the rate of these

processes is very small as compared with that of oxygen [Mitra, 1952].) After the subsidences, which predominate in the polar night mesosphere, the atomic oxygen produced by auroral activities recombines into molecular oxygen at lower altitudes, releasing the association energies.

It should be noted that, since the association (or recombination) is mostly due to three-body collisions in which the remaining particles carry away the excess recombination energy, the rate of this process is roughly proportional to the cube of the pressure. Consequently, atomic oxygen produced at high altitudes does not produce any significant effect on the warming of the lower mesosphere until it is advected into lower altitudes. In other words, part of the energy brought into the polar upper atmosphere by auroral particles as well as by the leakage particles from the radiation belts around the earth can be stored in the lower thermosphere in the form of dissociated oxygen. This energy is released in warming the upper mesosphere with prevailing subsidences in winter.

ENERGY FLUX OF AURORAL PARTICLES IN THE UPPER ATMOSPHERE

The magnetic field in the polar regions may be regarded as uniform and perpendicular to the earth's surface. In these regions charged particles impinging upon the atmosphere follow a spiral orbit along magnetic lines of force. The direct evidence for spiral motion of auroral protons is shown by Meinel [1951].

Since the rate of energy loss of these particles is the same in all directions, the pitch angle of a particle is constant throughout its range. In other words, incident particles maintain spiral motion with constant pitch angles, while the radius of gyration decreases as the particle loses energy at increasing atmospheric depths [Rhodes, 1955]. With regard to the constancy of pitch angle of the spiraling particles through the atmosphere, Singer [1958] has suggested an extension of Liouville's theorem to estimate the intensity of the nonrelativistic particles in the upper atmosphere. However, the application of Liouville's theorem can be extended only to the constant ionization loss, but not to the particles whose

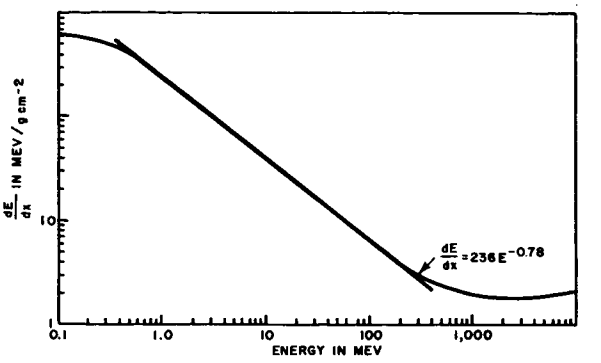


FIGURE 1.—Rate of energy loss of protons in air (below the minimum ionization loss) as a function of energy.

rates of energy loss change with their energy [Maeda, 1962].

The intensity of auroral protons whose energies are less than that of minimum ionization loss can be estimated by means of the so-called gross transformation [Rossi, 1952], which is, however, modified as shown in the following sections. On the other hand, the same method cannot be used for auroral electrons because scattering is predominant for low-energy electrons.

Low-Energy Protons

As can be seen in figure 1, the rate of energy loss E of protons per unit depth of air R can be approximated, below a certain energy value E_u, by the equation

$$-dE/dR = AE^{-m} \text{ Mev}/(\text{g cm}^{-2}) \qquad (1.1)$$

with the experimental data given by Segre [1953], where A and m are given in table 1 [Maeda and Singer, 1961] for E in Mev. There is also a lower limit of energy, E_l, for (1.1) to be valid, but the contribution below this limit to the total range is negligible.

TABLE 1.—*Energy Loss Coefficients in Air*

Particle	A	m	E_u	E_l
Proton*	236	0.78	300 Mev	400 kev
Electron†	0.38	0.96	100 kev	4 kev

* Based on Segre's [1953] data.
† Based on Katz and Penfold [1952].

Several direct observations show that the energy spectrum of auroral protons can be expressed by

$$j(E)\,dE=\alpha E^{-n}\,dE \qquad n>0 \qquad (3.2)$$

where $n\cong 5.5$ according to direct measurements by rocket-borne detector [McIlwain, 1960].

Since a simple relationship exists between the range of a particle at an atmospheric depth x (in g cm^{-2}), and the energy of the particle at another depth x', it is convenient to convert the energy spectrum into a range spectrum. It follows from (1.1) that

$$R=\int_0^E \frac{dE'}{AE'^{-m}} \qquad (1.3)$$

Combining (1.1), (1.2), and (1.3),

$$j(R)\,dR=kR^{-l}\,dR \qquad (1.4)$$

where

$$k=\alpha A^{-(n-1/m+1)}(m+1)^{-l}$$

and

$$l=(n+m)/(m+1)$$

Total Intensity of Spiraling Protons

If the total flux of auroral protons at the top of the atmosphere can be assumed to be isotropic, then it is given by

$$I_0=2\pi\int_{R_0}^{R_m} j(R)\,dR \qquad (1.5)$$

where R_m is the residual range corresponding to the maximum energy of incident protons, and R_0 is the threshold range set by the detector. We can assume that $R_m=\infty$ for $l>1$, so that (1.5) becomes

$$I_0=(2\pi k/l-1)\,R_0^{-(l-1)} \qquad (1.6)$$

The directional intensity of the flux at a depth x and at a pitch angle α is

$$j(x,\alpha)=\int_{(R_0+y)}^{R_m} kR^{-l}\,dR \qquad (1.7)$$

where $y=x/\cos\alpha$, the actual path length traversed by the protons in reaching the depth x.

The total flux, or the total integral intensity $I(x)$, at the depth x is

$$I(x)=2\pi\int_0^{\pi/2} j(x,\alpha)\,\sin\alpha\,d\alpha$$

$$=2\pi x\int_x^{\infty} j(y)\,(dy/y^2) \qquad (1.8)$$

From (1.5), (1.7), and (1.8), we finally obtain

$$I(x)/I_0=1-F_{l-1}(\xi) \qquad (1.9)$$

where $\xi=x/R_0$

$$F_0(\xi)=\frac{\ln(1+\xi)+\xi\ln(1+\xi^{-1})}{\ln R_m/R_0} \quad \text{for} \quad l=1$$

and

$$F_{l-1}(\xi)=(l-1)\xi\ln(1+\xi^{-1})-\xi\sum_{r=2}^{l-1}[(-1)^r/(r-2)]$$
$$\times(l-{}_1C_r)[1-(1+\xi^{-1})^{-(r-1)}] \quad \text{for} \quad l>1$$

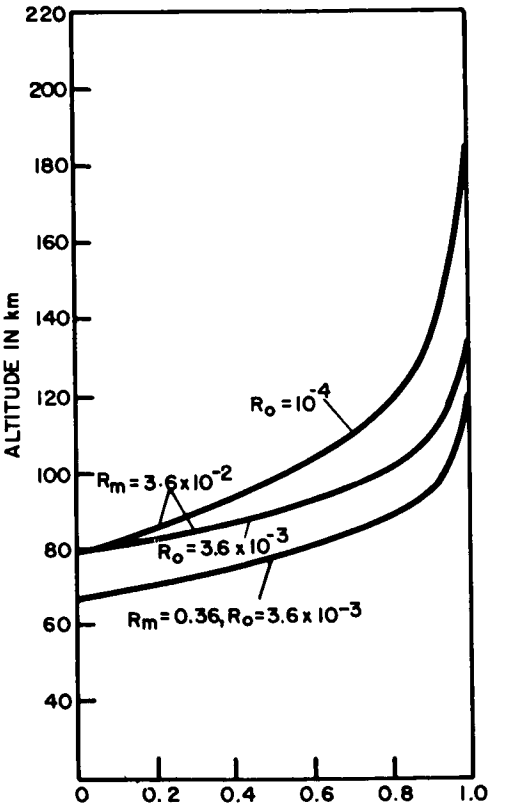


FIGURE 2.—Relative total intensity of spiraling protons $I(z)/I_0$ vs. altitude z for the case $l=1$, where the residual ranges 0.36, 3.6×10^{-2}, 3.6×10^{-3}, and 10^{-4} (in g/cm^{-2}) correspond to the energies 12, 5, 1.2, and 0.065 (in Mev), respectively.

In the last expression, R_m is taken as ∞, and

$$l-{}_1C_r=[(l-1)!/(l-1-r)!r!]$$

In figure 2, the relative total intensity, given by (1.9) with $l=1$, is plotted against altitude z, and in figure 3 similar curves for $l=1$, 2, and 3 are shown. From figure 2 we can see that the spiraling flux becomes zero below a certain depth which corresponds to the upper limit of the range spectrum R_m for vertical incidence. On the other hand, as is shown in figure 3, there is no sharp decrease of total flux for $R_m=\infty$. It should be noted that even for the same cutoff and upper limit of spectrums, the rate of decrease of total intensity with atmospheric depth becomes steeper for the steeper spectrum.

In these figures, the expression $I(x)/I_0$ is transformed into $I(z)/I_0$ (where z is the altitude of the level of atmospheric depth x) by making use of the relation between the atmospheric depth x (g/cm^2) and the altitude z (in kilometers). For this purpose, the following atmospheric model was used:

$z\leq 100$ km, Rocket Panel [1952] data.
$110<z\leq 400$ km, Kallmann's [1959] table.
$z>400$ km, King-Hele's [1962] graph.

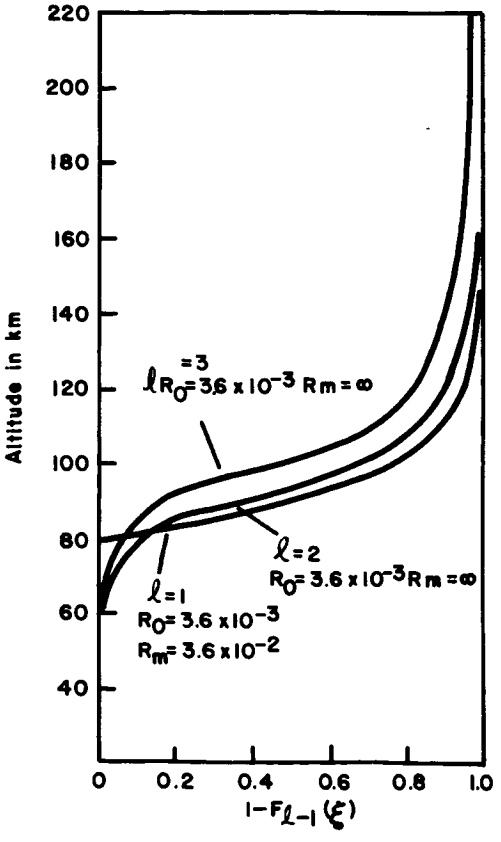


FIGURE 3.—Relative total intensity of spiraling protons $I(z)/I_0$ as a function of altitude z for the cases $l=1$, 2, and 3.

Omnidirectional Differential Spectrums

Since the rate of energy loss of protons penetrating into the atmosphere is larger for protons of lower energy (below minimum ionization), the flux of low-energy protons decreases faster than that of high-energy protons. Therefore, the form of the differential energy spectrum changes with the atmospheric depth. To illustrate this, the change of range spectrum of low-energy protons penetrating the atmosphere is computed.

The omnidirectional differential range spectrum at the top of the atmosphere is given by

$$i_0(R)\ dR=2\pi j(R)\ dR$$

and, at the depth x,

$$i_x(R)\ dR=2\pi\int_0^{\pi/2} j\left(R+\frac{x}{\cos\alpha}\right)\sin\alpha\ d\alpha$$

For the case of $j(R)\ dR=kR^{-l}\ dR$, setting $\cos\alpha=\mu$, and $x/R=\eta$,

$$i_x(R)\ dR=2\pi kR^{-l}\,dR\int_0^1[1+(\eta/\mu)]^{-l}\,d\mu \quad (1.10)$$

Using the range-energy relation for protons and the x-z curves, the omnidirectional differential energy spectrums $i_z(E)$ are shown for the case of $l=2$ in figure 4. From this, the effective power of the spectrum is seen to decrease with the atmospheric depth, i.e., with the altitude z.

Auroral Electrons

Although incoming protons with energies around several hundred kev are indicated by the observation of Doppler-shifted Balmer lines in the auroral spectrum, it is found that the major fraction of auroral particles in weak auroras consist of electrons with energies below 50 kev [McIlwain, 1960]. The integral energy spectrum is about

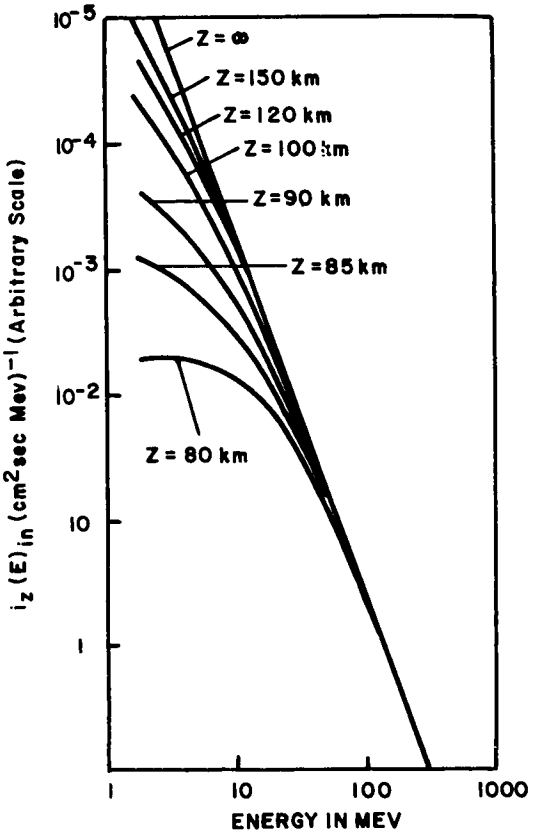


FIGURE 4.—Omnidirectional differential energy spectrum of spiraling protons $i_z(E)$ at several altitudes z in kilometers.

$4\times10^{10}\ E^{-3}/\text{cm}^2$ sec ster, where E is expressed in kev. [Originally, the observed data were approximated by $2.5\times10^9 \exp(-E/5)$, (E in kev) for the energy range between 3 and 30 kev. In this energy interval, both expressions agree within the accuracy of the measurements.]

Particle Range of Electrons in Air

The rate of energy loss, $-dE/dR$, is given theoretically by Bethe [1933], and the residual range is obtained by

$$R=\int_0^E \frac{dE}{(-dE/dR)}$$

Because of multiple Coulomb scattering and straggling effects, both of which are predominant for low-energy electrons traveling through the air, the rate of energy loss given by Bethe's formula is inadequate for calculating the residual range of low-energy electrons.

On the other hand, the practical range R_p, which is the projected depth of the incident electrons in the initial direction, has been defined and is experimentally measurable [Katz and Penfold, 1952]. Since the theoretical range R is the total length measured along the tortuous path of an electron in the air, the practical range R_p is always shorter than R. A comparison of the rate of energy loss in terms of R_p and of R is shown in figure 5. From this figure, it is apparent that the difference between dE/dR_p and dE/dR increases with decreasing electron energy, because multiple scattering increases at low energies. (The dE/dR curve in figure 5 is taken from Fermi's text [Orear, Rosenfeld, and Schluter, 1950].) The coefficients A and m in table 1 indicate an empirical approximation of $-dE/dR_p$ by (1.1), E being given in Mev.

The angular distribution of the scattered electrons in multiple Coulomb scattering is known to be approximately Gaussian. However, if the penetration depth of electrons is far larger than the mean free path of Coulomb scattering of electrons in the air, λ_c, the distribution is no longer Gaussian, but is of the form $\cos^2\theta$, where θ is the scattering angle [Wu, 1960].

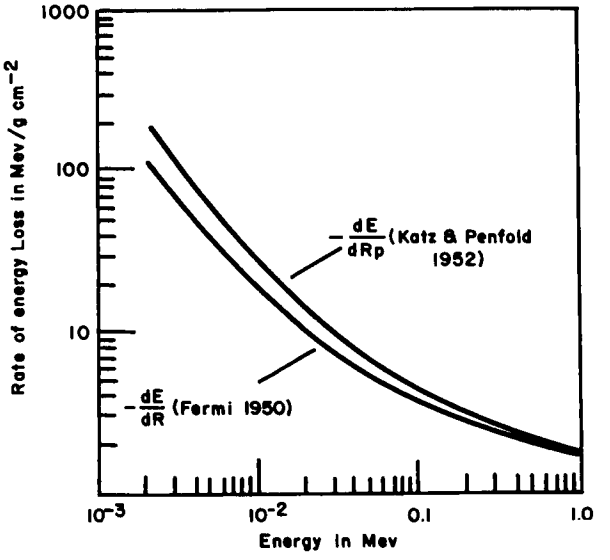


FIGURE 5.—Rate of energy loss of electrons in air in terms of the true range R and of the practical range R_p as a function of electron energy.

The corresponding mean angle of scattering is around 35° and remains the same with further increase of thickness. In other words, the penetration of electron flux deeper than λ_c in the air can almost be regarded as diffusion. In the diffusion of electrons in the atmosphere, there is another factor called 'back diffusion thickness,' x_d, beyond which the backscattered electron is saturated. If the penetration depth of electrons exceeds x_d, complete diffusion of electrons starts. However, in air the ratio of the number of backscattered electrons to that of incident primary electrons at saturation, i.e., the back diffusion coefficient, is less than 0.2.

In figure 6, λ_c, x_d, and R_p are given as functions of the incident electron energy. λ_c is computed by

$$\lambda_c = A_a/\sigma_c N$$

where A_a is the effective atomic weight of air nuclei, 14.78; N is Avogadro's number, 6.024×10^{23}; and σ_c is the total cross section for Coulomb scattering. x_d is estimated from experimental data on aluminum, multiplied by a reducing factor 0.34, which is approximately the ratio of the total cross section of Coulomb scattering in air to that in aluminum.

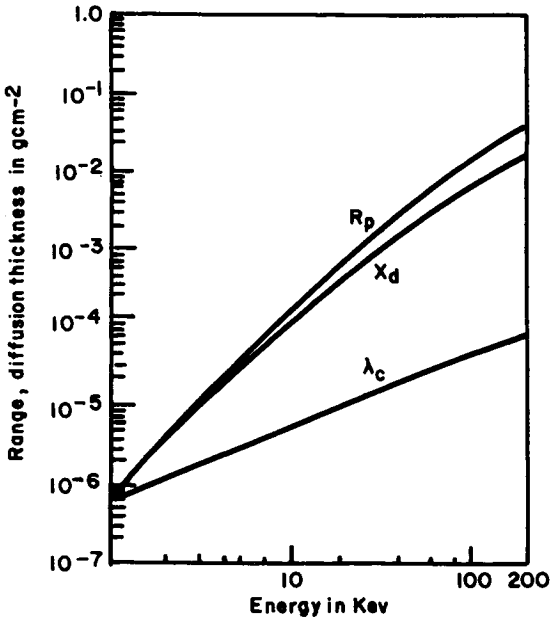


FIGURE 6.—Practical range R_p, back diffusion thickness x_d, and the mean free path for Coulomb scattering λ_c, of electrons in air as functions of electron energy E in Kev.

Absorption of Electrons in Air

As shown in the previous section, the penetration of electrons in air beyond the depth of several times λ_c can be regarded approximately as diffusion. This is due to multiple Coulomb scattering and the straggling effect caused by frequent large fractional energy transfers in the air. Thus, even if the incident electron flux is monoenergetic, the energy of the electrons at a certain depth is no longer monochromatic but spread widely toward the lower energies; and the electrons of lowest energy are captured by air nuclei. Consequently, the intensity of auroral electrons in the atmosphere decreases gradually with atmospheric depth.

The rate of decrease of electron flux per unit atmospheric depth is a function of the energy of the incident electrons. This has been observed in several laboratory experiments [Lenard, 1903; Becker, 1905; Crowther, 1906; Schmidt, 1909; Mayer, 1914; and Friman, 1916]. According to these measurements, the differential intensity of electrons with initial energy E, at the depth x (in grams per centimeter^2) is approximately

$$j(E, x)\, dE = j_0(E)\, \exp[-x/\lambda(E)]\, dE$$

where $j_0(E) = j(E, 0)$, the initial differential intensity of electrons with a kinetic energy of E, and $\lambda(E)$, which can be called the 'attenuation mean free path of electrons in the air,' is the inverse of the mass absorption coefficient of electrons with initial energy E.

Figure 7 shows $\lambda(E)$ as a function of the initial electron energy, derived from the experimental results mentioned above, which are also summarized in a table by Bethe [1933]. It is quite evident from figure 7 that the attenuation mean free path of electrons with initial energy E can be expressed empirically by

$$\lambda(E) = bE^{m'} \text{ g/cm}^2 \qquad (1.11)$$

where the numerical values of the constants b and m' are given in table 2.

Although the power m' decreases significantly above about 300 kev, from 2.2 to 1.33, $m' = 2.2$ can be used for all energy ranges above 5 kev.

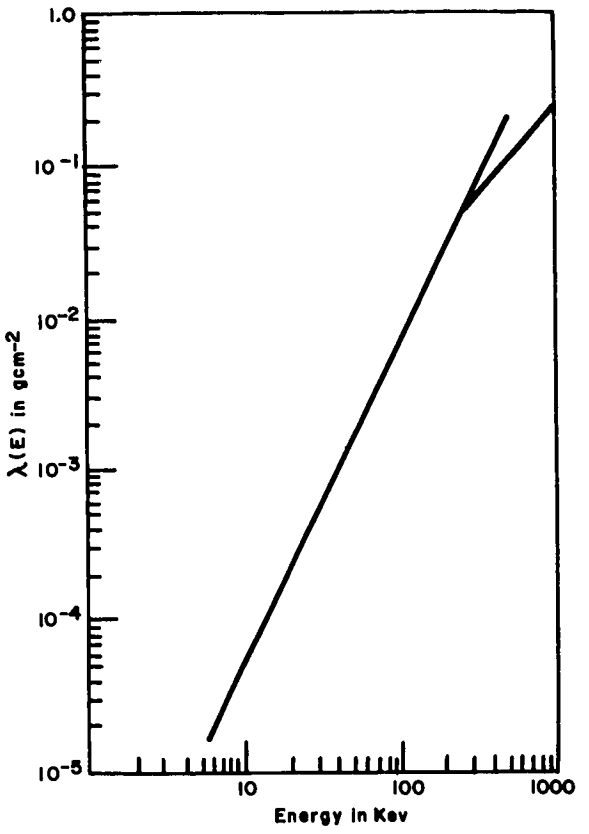


FIGURE 7.—Attenuation mean free path of electrons in air as a function of incident electron energy.

Since the intensity of high-energy electrons is very small, owing to the sharp drop of the energy spectrum on the high-energy side, the use of $m'=2.2$ in (1.11) beyond 300 kev does not introduce any serious errors in the present calculation.

TABLE 2.—*Constants of Attenuation Mean Free Path of Electron in Air, E*

b	m'	Energy Range
3.15×10^{-7}, E in kev 1.26, E in Mev	2.2	5 kev $<E<$ 300 kev
2.94×10^{-5}, E in kev 0.29, E in Mev	1.33	0.3 Mev $<E<$ 3 Mev

Total Intensity of Auroral Electrons

The differential intensity of electrons entering the atmosphere having pitch angle α with respect to the vertical magnetic field at the atmospheric depth x can be written

$$j(E,x,\alpha)\,dE=j_0(E)\exp(-x/\lambda(E)\cos\alpha) \quad (1.12)$$

where $j_0(E)$ is the differential energy spectrum of incident electrons and $\lambda(E)$, the attenuation mean free path of electrons, is given by the empirical formula 1.11. Thus, the total flux of electrons at the atmospheric depth x is

$$I(x)=2\pi\int_0^{\pi/2}\int_{E_0}^{\infty} j_0(E)\cdot\exp\left[-\frac{\chi}{\lambda(E)\cos\alpha}\right]\cdot\sin\alpha\,d\alpha\,dE \quad (1.13)$$

The initial differential energy spectrum $j_0(E)$ is assumed to be given by (1.2), where $n\cong4$, according to the direct measurements by McIlwain [1960]. Using the substitutions, $y=x/\cos\alpha$ and $t=1/\lambda(E)$, (1.13) becomes

$$I(x)=2\pi x\int_{E_0}^{\infty} j_0(E)\int_{\chi}^{\infty}\exp[-y/\lambda(E)]y^{-2}\,dy\,dE$$

$$=\frac{2\pi\alpha}{m'}b^{\beta}\cdot\int_0^{1/\lambda_0} t^{\beta-1}G(tx)\,dt$$

where

$$\beta=(n-1)/m' \qquad \lambda_0=\lambda(E_0)=bE_0^{-m'}$$

and $G(u)$ is the so-called Gold integral [Rossi, 1952], which is defined by

$$G(u)=\int_1^{\infty}\frac{e^{-us}}{s^2}\,ds=u\int_u^{\infty}\frac{e^{-s}}{s^2}\,ds$$

The initial total intensity of auroral electrons is given by letting $x=0$ in (1.13)

$$I_0=(2\pi\alpha/m')b^{\beta}(\lambda_0^{-\beta}/\beta)$$

Therefore, the relative total intensity at the

depth x is

$$\frac{I(x)}{I_0}=\beta\lambda_0{}^{\beta}\int_0^{1/\lambda_0} t^{\beta-1}G(tx)\,dt \qquad (1.14)$$

The two curves in Figure 8 represent $I(x)/I_0$ for $n=4(\beta=1.4)$ and $n=3.2(\beta=1)$ with $E_0=10$ kev. By comparing figures 8 and 2, it can be seen that the decrease of electron flux in the upper atmosphere is not as steep as that of the protons in both cases for the value of n.

Omnidirectional Differential Spectrums

Since the attenuation of the electron flux is larger for electrons of lower energies, the energy spectrum of electrons changes with atmospheric depth. To see this variation, the omnidirectional

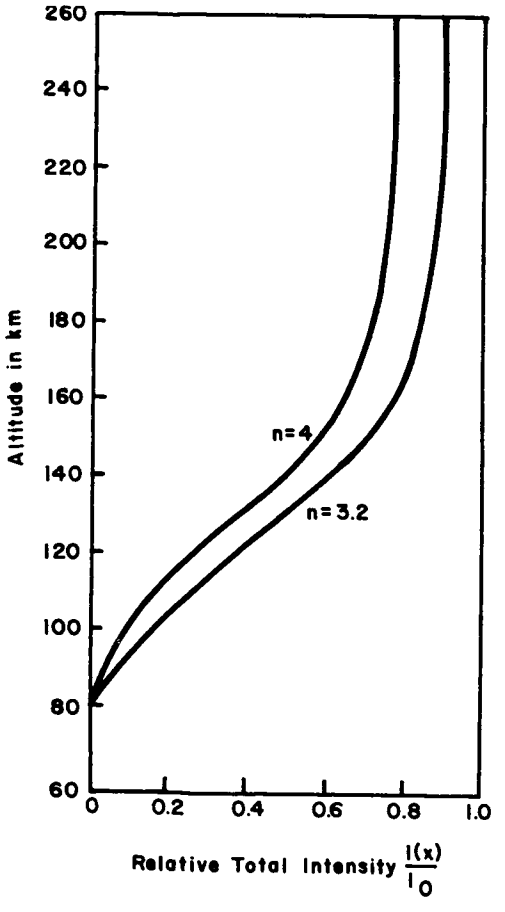


FIGURE 8.—Relative total intensities of spiraling electrons for two differential energy spectrums as functions of altitude.

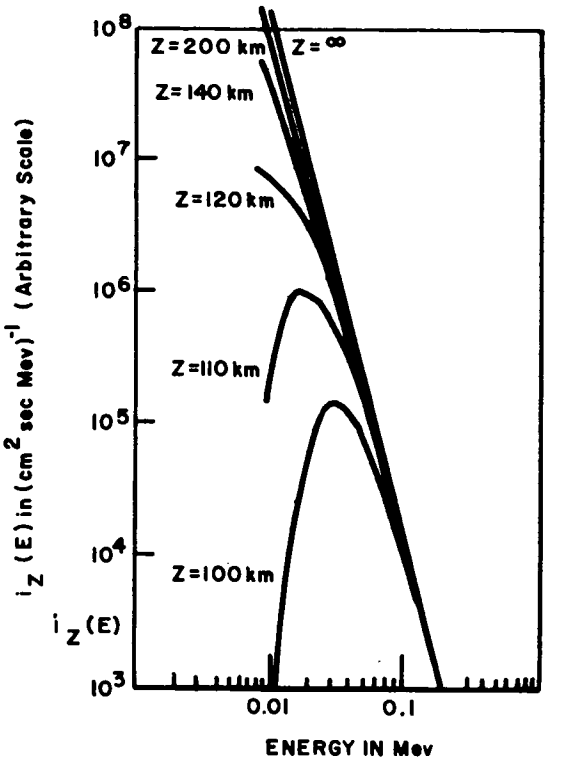


FIGURE 9.—Omnidirectional differential spectrums of auroral electrons with power spectrum $n=4$.

differential energy spectrums of electrons at the atmospheric depth x can be calculated by

$$i_x(E)\,dE=2\pi\int_0^{\pi/2} J(E)\cdot\exp\left[\frac{x}{-\lambda(E)}\cos\alpha\right]\sin\alpha\,d\alpha\,dE$$

$$=2\pi\alpha E^{-n}x\int_x^{\infty} y^{-2}\exp[-y/\lambda(E)]\,dy\,dE$$

Since the omnidirectional differential intensity at the top of the atmosphere is given by

$$i_0(E)\,dE=2\pi j_0(E)\,dE$$

The omnidirectional differential intensity of electrons at the atmospheric depth x is given by

$$i_x(E)=i_0(E)g[x/\lambda(E)] \qquad (1.15)$$

where $G(u)$ is Gold integral as in the previous section.

By making use of the x–z curve, $i_x(E)$'s are easily transformed into the $i_z(E)$'s shown in figure 9, from which the intensity of the low-energy side of the spectrum is seen to decrease very rapidly. Owing to this rapid decrease, the low-energy spectrum becomes a round peak with increasing atmospheric depth, and the position of this peak shifts toward the high-energy side with increasing atmospheric depth.

RATE OF DISSOCIATION OF MOLECULAR OXYGEN BY AURORAL PARTICLES

According to the Frank–Condon principle, oxygen molecules will dissociate upon the impact of charged particles or photons having kinetic energies higher than 5.08 ev. This value corresponds to the transition energy threshold from the ground level $^3\Sigma_g^-$ to the lowest repulsive level $^3\Sigma_u^-$.

The total cross section for the dissociation of molecular oxygen by electrons was measured by Glocker and Wilson [1932] and estimated by Massey [1938] and Mott and Massey [1949]. By combining these results, the total cross section is obtained as a function of the energy of impact electrons. This cross section increases to a maximum of 8 ev, and declines at about 30 ev as the inverse of E (figure 10). (The finite cross section below 5 ev can be explained by the dissociation through excitations of oxygen molecules by electron impacts.) The maximum value of this cross section is of the order of 10^{-18} cm^2, which is of the same order as the molecular cross section estimated from the absorption coefficient of the Schumann–Runge band (1750–1400 A; 6.7–9.7 ev). It is interesting to note that the efficiencies of O_2 dissociation by electrons and photons around the energy levels of 10 ev are both of the same order.

On the other hand, the dissociation of molecular osygen by the direct impact of primary protons is not important unless their energy is of the order of 10 ev, since the effective energy for dissociation is below 30 ev. Thus, the effective particles for O_2 dissociation are not composed of primary particles, but mainly of secondary and tertiary particles produced in the air by the incoming primary protons and electrons.

In this respect the total specific ionization $j_t(E)$ is a most convenient quantity since it is the total number of ion pairs per g/cm^2 produced by primary particles of energy E (in ev) and by all secondary electrons [Rossi, 1952, p. 45]. The total specific ionization is

$$j_t(E) = k(E)/V_0 \tag{2.1}$$

where $k(E)$ corresponds to the primary ionization given by $-dE/dx$ (or $-dE/dR$) and V_0, the average energy expended per ion pair produced, is determined experimentally, i.e., 32 ev for electrons in air and 36 ev for protons in air.

The rate of O_2 dissociation $r_p(z)\ dz$ at the altitude z, by auroral protons is

$$r_p(z)\ dz = \int_{E_m}^{\infty} N(z)\left[\int_{W_0}^{E} \sigma(W) j_{t,p}(W, E, z)\ dW\right] \cdot i_p(E, z)\ dE\ dz \tag{2.2}$$

where

$N(z)$ is the number density (cm^{-3}) of molecular oxygen.

$\sigma(W)$ is the differential cross section (cm^2) for O_2 dissociation by electrons of energy W, as shown in figure 10.

$j_{t,p}(W, E, z)$ is the number (g/cm^2)$^{-1}$ of secondary electrons of energy W produced by protons of energy E per unit thickness of air at the altitude z.

$i_p(E, z)$ is the omnidirectional differential intensity of auroral protons at the altitude z (in kilometers), which corresponds to the atmospheric depth x (in g/cm^2). This is given by (1.10) and shown in figure 4. [(1.10) shows the differential range spectrum at the atmospheric depth x, $i_x(R)$, which is, however, easily transformed into $i_z(E)$ by making use of x–z curve, i.e.,

$$i_z(E) = i_z(R)\ (dR/dE) = i_x(R)\ (dR/dE)$$

where dR/dE is shown in figure 1.]

W_0 is the threshold energy for O_2 dissociation.

E_m is the lowest energy of primary protons.

The number density of molecular oxygen is

$$N(z) = 2.16 \times 10^{22} \rho(z)\, C(z) \tag{2.3}$$

where $\rho(z)$ is air density at the altitude z in g/cm^3, $C(z)$ is the fractional composition of O_2 by weight.

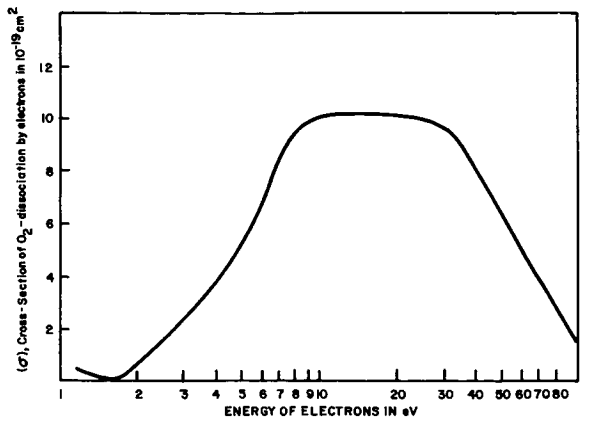


FIGURE 10.—Total cross section of O_2 dissociation by electrons as a function of kinetic energy of impact electrons.

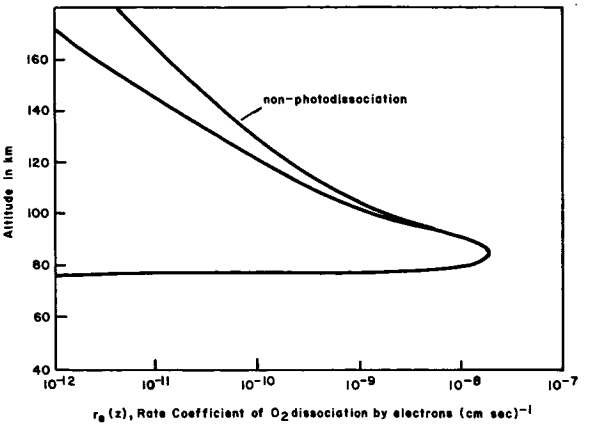


FIGURE 12.—Rate coefficients for O_2 dissociation by auroral electrons with differential spectrum $E^{-4}dE$, as a function of altitude.

In the integrand of (2.2), the terms inside the brackets can be written as

$$\int_{W_0}^{E} \sigma(W) j_{t,p}(W, E, z)\, dW \simeq \frac{\sigma_{\text{int}}}{W_e} \frac{k_p(E)}{V_{e,p}} \rho(z)$$

where

$$\sigma_{\text{int}} = \int_{W_0}^{\infty} (W)\, dW \simeq 6 \times 10^{-17} \text{ cm}^2 \text{ ev}$$

is the integral cross section for O_2 dissociation by electrons (see figure 10). W_e is the effective energy for O_2 dissociation, about 20 ev, $\rho(z)$ is the air density at the altitude z, (g/cm^3), and $k_p(E)/V_{0,p}$ is the total specific ionization of a proton with energy E in the air, i.e., the number

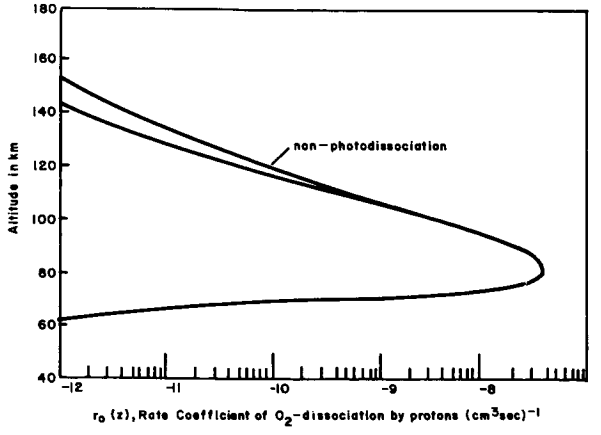


FIGURE 11.—Rate coefficient for O_2 dissociation by auroral protons with differential spectrum $E^{-2.8}dE$, as a function of altitude.

of secondary electrons per unit depth of air in (g/cm^2)$^{-1}$, where $V_{0,p} = 36$ ev.

Evaluating the foregoing equation, we have

$$\int_{W_0}^{E} \sigma(W) j_{t,p}(W, E, z)\, dW \simeq 8.33 \times 10^{-20} \rho(z) k_p(E) \quad (2.4)$$

Substituting (2.3) and (2.4) into (2.2), we finally obtain

$$r_p(z)\, dz \simeq 1.8 \times 10^3 \rho^2(z) C(z) \int_{E_m}^{\infty} k_p(E) i_p(E, z)\, dE\, dz \quad (2.5)$$

where $k_p(E)$ is shown in figure 1.

Similarly, the rate of O_2 dissociation due to auroral electrons at the level z is given by

$$r_e(z)\, dz \simeq 2.0 \times 10^3 \rho^2(z) C(z) \cdot \int_{E_m}^{\infty} k_e(E) i_e(E, z)\, dE\, dz \quad (2.6)$$

where $k_e(E)$ is shown in figure 5 by $-dE/dR_p$, and $i_e(E, z)$ is shown in figure 9 for the case of the incident energy spectrum $aE^{-4}\, dE$.

It should be noted that the square of the atmospheric density in (2.5) and (2.6) is due, first, to the target distribution (the O_2 distribution) in the atmosphere given by (2.3) and, second, to the reduction of the rate of ionization loss of primary particles from ev/(g/cm^2) to ev/cm.

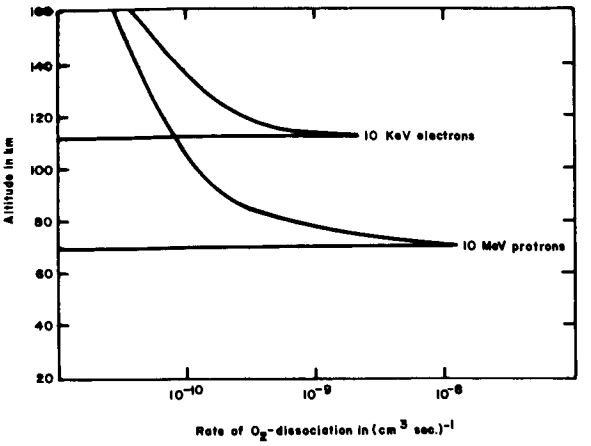


FIGURE 13.—Rates of O_2 dissociation by monoenergetic electrons (10 kev) and protons (10 Mev) in the upper atmosphere, normalized for initial total intensity, i.e., rate coefficients as functions of altitude.

The values $r_p(z)$ and $r_e(z)$, normalized to the incident total intensity, are shown in figures 11 and 12, respectively.

For these numerical calculations the two extreme cases of $C(z)$ are assumed: (1) The minimum distribution of O_2, where some part of the atmospheric molecular oxygen is dissociated by vertical insolation. Specifically, the fractional composition given by Johnson [1962] is used. (2) The maximum distribution of O_2, where no photodissociation of oxygen is assumed and simply constant relative concentration of molecular oxygen, $C=0.23$, is extended to the top of atmosphere. The effect of photodissociation on the auroral dissociation is clearly negligible below the mesopause (about 86 km), where the dissociation by auroral particles is most effective.

CONCLUSIONS

To calculate the final atomic oxygen distribution due to O_2 dissociation by auroral particles we must assume the duration of aurora, total flux of incident particles or total energy flux, and the rate of association at different altitudes. Instead of presenting cases with several combinations of these factors, we make the following points:

1. The rate coefficient of O_2 dissociation due to auroral protons and electrons is almost the same as that due to the solar ultraviolet radiation, which is known as the Schumann–Runge continuum (1400–1750 A); i.e., it is of the order of 10^{-18}/cm^3 sec at the maximum.

2. Although the profile of the atomic oxygen production rate in the atmosphere due to auroral particles depends on the energy spectrum, it is quite similar to that of photodissociation due to solar ultraviolet radiation. The maximum height of O_2 dissociation is roughly around 90 km in both cases [Friedman, 1960; Hinteregger, 1961].

If the incident particles are monoenergetic, the lower boundary of the profile becomes very steep, as it is shown in figure 13, where the dissociation rates for 10-kev electrons and for 10-Mev protons are plotted. These lower boundaries, given in terms of the penetration depth of spiraling particles, are drawn in figure 14 by using the atmospheric depth-altitude relation (x–z curve) and the the range-energy relation for protons in the air [Segre, 1953].

According to direct measurements by rocket-borne instruments, the energy flux of auroral particles is roughly 20 ergs/cm^2 sec at weak

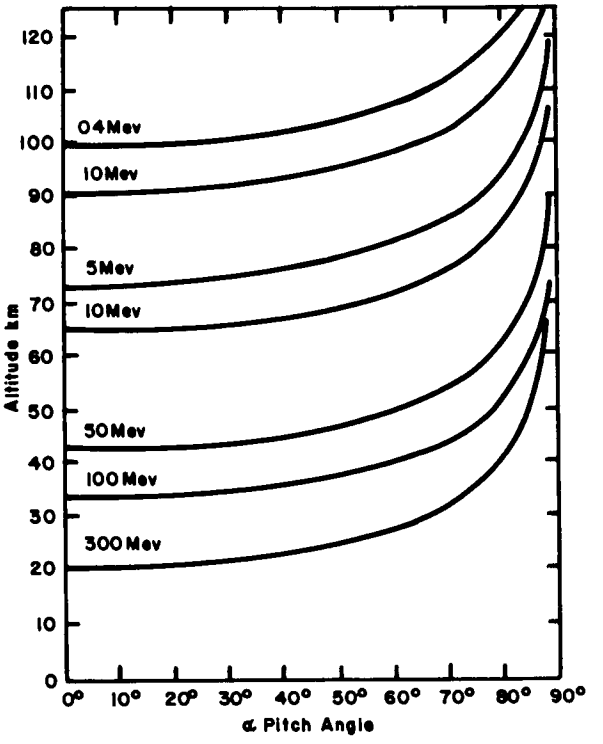


FIGURE 14.—Penetration depth of spiraling monoenergetic protons in the atmosphere as a function of the pitch angle.

auroras and exceeds 10^3 ergs/cm^2 sec at severe bright auroras [McIlwain, 1960]. These values are also consistent with the estimation made by Chamberlain [1961] based on auroral luminosities. On the other hand, the average energy flux of solar ultraviolet around the Schumann–Runge continuum is approximately 300 ergs/cm^2 sec or less [Friedman, 1960; Hinteregger, 1961].

Since the rate of association of atomic oxygens is smaller at higher altitudes, the maximum concentration of atomic oxygen is at a higher altitude than the production peak of oxygen atoms. According to Nicolet [1960] the maximum concentration of oxygen atoms under equilibrium conditions for the dissociation rate of 10^4/cm^3 sec is of the order of 10^{11} atoms/cm^3 at the 110-km level. It should be noted that during severe aurora, in which the flux of electrons is of the order of 10^{12}/cm^2 sec [McIlwain, 1960], the rate of production of oxygen atoms due to auroral electrons is of the order of 10^4 atoms/cm^3 sec at the level of maximum production (around 85 km).

It has been shown that if there is no strong downward advection above the mesosphere, the lifetime of atomic oxygen around the height of maximum concentration (>100 km) is more than 4 months, but at the altitude of the production peak (about 85 km), it is less than one month [Nicolet and Mange, 1954]. Furthermore, if there is a strong subsidence as in the case of polar mesosphere in winter, this lifetime of oxygen atoms should be shortened drastically. In this respect, the production of oxygen atoms by auroral particles in the polar upper atmosphere during the winter might be one of the most important processes in supplying atomic oxygen which in turn would provide the largest contribution to the heating of the polar night mesosphere as well as the cooling of the thermosphere [Maeda, 1962].

Finally, it should be emphasized that direct measurement of atomic oxygen in the upper atmosphere, which has not been possible so far because of experimental difficulties, is very desirable in order to clarify the mysterious warm polar night mesosphere and the irregular warmings of middle-latitude mesospheres.

REFERENCES

Becker, A., Messungen an Kathodenstrahlen, *Ann. Phys.*, [4]**17,** 381–470, 1905.

Bethe, H., Quantenmechanik den Ein- und Zwei-Elektronon-Probleme, *Handbuch der Physik*, 24, part 1, pp. 273–560, Springer-Verlag, Berlin, 1933.

Bothe, W., Durchgang von Elektronen durch Materie, *Handbuch der Physik*, **22,** part 2, pp. 1–74, Springer-Verlag, Berlin, 1933.

Chamberlain, J. W., *Physics of the Aurora and Airglow*, Academic Press, New York, International Geophysics Series 2, 1961.

Crowther, J. A., On the coefficient of absorption of the beta-rays from uranium, *London, Edinburgh, and Dublin Phil. Mag. and J. of Sci.*, [6]**12,** 379–392, 1906.

Friedman, H., The sun's ionizing radiations, in *Physics of the Upper Atmosphere*, edited by J. A. Ratcliffe, pp. 133–218, Academic Press, New York, 1960.

Friman, E., Über die Absorption und Diffusion schneller Kathodenstrahlen (β-Strahlen) in Gasen und Dampfen, *Ann. Phys.*, [4]**49,** 373–418, 1916.

Glockler, G., and J. L. Wilson, The activation of molecular oxygen by electron impact, J. Am. *Chem. Soc.* **54,** 4544–4558, 1932.

Hinteregger, H. E., Preliminary data on solar extreme ultraviolet radiation in the upper atmosphere, *J. Geophys. Res.* **66,** 2367–2380, 1961.

Ishikawa, G., Solar corpuscular radiation as a heat source of the upper atmosphere, *Papers in Meterol. Geophys. Tokyo*, **10,** 93–123, 1959.

Jastrow, R., Outer atmospheres of the earth and planets, *J. Geophys. Res.*, **64,** 1789–1798, 1959.

Johnson, F., The physical properties of the earth's *Progress in the Astronautical Sciences*, edited by S. F. Singer, chapter 2, North-Holland Publishing Co., Amsterdam, 1962.

Kallmann, H. K., A preliminary model atmosphere based on rocket and satellite data, *J. Geophys. Res.*, **64,** 615–623, 1959.

Katz, L., and A. S. Penfold, Range-energy relations for electrons and the determination of beta-ray end-point energies by absorption, *Rev. Mod. Phys.* **24,** 28–44, 1952.

Kellogg, W. W., Chemical heating above the polar mesosphere in winter, *J. Meteorol.*, **18,** 373–381, 1961.

Kellogg, W. W., Warming of the polar mesosphere and lower ionosphere in winter, *Rand Corp. Tech. Rept. P*-2032, August 1960.

King-Hele, D. G., Properties of the atmosphere revealed by satellite orbits, in *Progress in Astronautical Science*, vol. 1, edited by S. F. Singer, North-Holland Publishing Co., Amsterdam, 1962.

Krassovsky, V. I., Energy source of the upper atmosphere, *Planetary Space Sci.*, **1**(1), 14–19, 1959.

Lenard, P., Über die Absorption von Kathodenstrahlen vershiedener Geschwindigkeit, *Ann. Phys.*, [4]**12,** 714–744, 1903.

MAEDA, K., On the heating of the polar upper atmosphere, *NASA Tech. Rept. TR-R-141*, 1962.

MAEDA, K., and S. F. SINGER, Energy dissipation of spiraling particles in the polar atmosphere, *Arkiv Geofysik*, **3**(21), 531–538, 1961.

MASSEY, H. S. W., *Negative Ions*, Cambridge Univ. Press, London, 1938.

MAYER, F., Über sekundäre Kathodendtrahlung in Gasen bei geringer Geschwindigkeit der Primarstrahlen und Über deren Absorption, *Ann. Phys.*, [4]**45,** 1–28, 1914.

MCILWAIN, C. E., Direct measurement of particles producing visible aurorae, Ph.D. Thesis, State Univ. of Iowa, June 1960.

MEINEL, A. B., Doppler-shifted auroral hydrogen emission, *Astrophys. J.*, **113,** 50–54, 1951.

MITRA, *The Upper Atmosphere*, 2nd ed., Asiatic Society, Calcutta, 1952.

MOTT, N. F., and H. S. W. MASSEY, *The Theory of Atomic Collisions*, 2nd ed., Clarendon Press, Oxford, 1949.

NICOLET, M., The properties and constitution of the upper atmosphere, in *Physics of the Upper Atmosphere*, edited by J. A. Ratcliffe, pp. 17–71, Academic Press, New York, 1960.

NICOLET, M., and P. MANGE, The dissociation of oxygen in the high atmosphere, *J. Geophys. Res.*, **59,** 15–45, 1954.

NORDBERG, W., and W. G. STROUD, Results of IGY rocket-grenade experiments to measure temperatures and winds above the island of Guam, *J. Geophys. Res.*, **66,** 455–464, 1961.

OREAR, J., A. H. ROSENFELD, and R. A. SCHLUTER, *Nuclear Physics, a Course Given by Enrico Fermi at the University of Chicago*, Univ. of Chicago Press, Chicago, 1950.

RHODES, R. M., Study of auroral particles, M.S. Thesis, Univ. of Maryland, 1955.

Rocket panel, pressures, densities, and temperatures in the upper atmosphere, *Phys. Rev.* **88,** 1027–1032, 1952.

ROSSI, B., *High Energy Particles*, Prentice-Hall, New York, 1952.

SCHMIDT, H. W., Beitrag zur Frage über den Durchgang der Beta-Strahlen durch Materie, *Phys. Z.*, **10,** 929–948, 1909.

SEGRE, E., editor, *Experimental Nuclear Physics*, John Wiley and Sons, New York, 1953.

SHEPPARD, P. A., Dynamics of the upper atmosphere, *J. Geophys. Res.*, **64,** 2116–2121, 1959.

SINGER, S. F., The primary cosmic radiation and its time variations, *Progr. Elem. Particle Cosmic Ray Phys.*, **4,** 203–336, 1958.

STROUD, W. G., W. NORDBERG, W. R. BANDEEN, F. L. BARTMAN, and P. TITUS, Rocket-grenade measurements of temperatures and winds in the mesosphere over Churchill, Canada, *J. Geophys. Res.*, **65,** 2307–2323, 1960.

WU, C. S., The interaction of beta particles with matter, in *Nuclear Spectroscopy*, part A, edited by Fay Ajzenberg-Selove, pp. 15–30, Academic Press, New York, 1960.

MASS SPECTROMETRIC INVESTIGATIONS OF THE ATMOSPHERE BETWEEN 100 AND 227 KM ABOVE WALLOPS ISLAND, VIRGINIA

EDITH MEADOWS-REED AND CHARLES R. SMITH

Goddard Space Flight Center

The diffusive separation of argon and nitrogen was measured by means of a Bennett radio frequency mass spectrometer carried on an Aerobee-150A rocket, NASA 4.14, launched at 1141 EST, November 15, 1960. The primary purpose of the spectrometer measurements was to obtain temperate latitude diffusive separation data to compare with those measurements obtained in the arctic (Fort Churchill, Canada) during the International Geophysical Year. The results indicated that diffusive separation occurred above about 111 km and that the altitude profile of the A/N_2 ratio is somewhat similar to that obtained during the two night flights in the arctic atmosphere.

INTRODUCTION

The flight of Aerobee-150A NASA 4.14 marked the 17th time that American scientists have flown the Bennett radio frequency mass spectrometer for the purpose of measuring the composition of the neutral particles of the upper atmosphere.

The first two flights of this instrument were carried out, using Aerobee rockets, by the Applied Physics Laboratory of Johns Hopkins University in 1950 and 1951.[1] The next 13 flights were made by the Naval Research Laboratory (NRL) from 1952 through 1959. In five of the NRL flights, rocket and instrumentation both performed as intended. However, the first of these five[2] was contaminated to a great extent by rocket gas; the designs of the experiment and the vehicle were modified accordingly. Four successful firings, during the period 1956–1958, were made in connection with the International Geophysical Year (IGY) program.[3]

All these experiments excepting the last one (in 1958) were designed to scan a mass range of about 8 to 48 atomic mass units (AMU) and were conducted primarily for the purpose of detecting diffusive separation of argon and molecular nitrogen. The last experiment, in 1958, scanned a mass range of 0.7 to 5 AMU and was an exploratory flight for the detection of hydrogen and helium.

Finally, as one of the initial efforts in the NASA program for studying the structure of planetary atmospheres, the Aerobee-150A rockets designated NASA 4.09 and 4.14 were instrumented with Bennett mass spectrometers patterned after those used by NRL in the IGY. Satisfactory data about the neutral composition were not obtained from the flight of NASA 4.09. The data obtained from NASA 4.14 are the principal subject of this report.

Scientists of the USSR have conducted four successful flights in which neutral composition was measured with the radio frequency mass spectrometer.[4,5] Their Model MX-6401 instrument had higher resolving power (1 part in 45) and a longer mass range (1 to 4 and 12 to 56 AMU) than did the American instruments. The Model-6403, an improved, lighter version of the MX-6401, is more sensitive by a factor of 10. The MX-6401 was flown twice in the daytime in July 1959 and also in the late afternoon in November 1961. The MX-6403 was flown at midnight in September 1960. In the 1959 and 1960 flights the mass spectrometer was housed in a cannister which was ejected from the vehicle. In the 1961 flight, the

mass spectrometer was not separated from the rocket, but was housed in a special container in the payload.

With the exception of the fourth USSR flight (November 1961) which reached an altitude of 430 km, all of the above mentioned flights, American and USSR, reached altitudes of less than 250 km.

Several other groups have flown other types of mass spectrometers to study the neutral gases. The first was a magnetic mass spectrometer prepared by R. C. Edwards and J. V. Fay, Jr.,* and flown on a V-2 in 1949.[1,6] Others include magnetic mass spectrometers (C. Y. Johnson, private communication), time-of-flight spectrometer (R. Narcissi, private communication), and the quadrupole "massenfilter"† (E. J. Schaefer, private communication). Rocket and instrumental difficulties have resulted in less than optimum performance of these experiments.

INSTRUMENTATION

Rockets NASA 4.09 and 4.14 were instrumented at Goddard Space Flight Center to measure, from 100 km to the peak of flight, the pressure, temperature, density, and composition of the neutral particles and positive ions in the earth's atmosphere. The instruments used were Bennett radio frequency mass spectrometers for neutral particles, similar mass spectrometers for atmospheric ions, and Bayard–Alpert ionization pressure gages. The mass spectrometers were patterned after those used by NRL during the IGY.[3,7] Although the technique of using ion gages was the same as that used by NRL during the IGY,[8] the first flight of the Westinghouse No. 4250 version of the Bayard–Alpert gage was aboard NASA 4.14. The locations of the various sensors are shown in figure 1. This instrumentation was not adequate to give all the aforementioned parameters with constant accuracy over the entire altitude range from 100 km to almost 250 km; but over certain portions of this range, it could provide very good data.

The NASA 4.09 flight was made at 1047 EST on 29 April, 1960, at Wallops Island, Virginia (latitude 37°50′06″ N, longitude 75°29′11″ W) to an altitude of 248 km. Only the ion mass spectrometer performed properly. The NASA 4.14 flight was made at 1141 EST, 15 November 1960, to 227.3 km, also from Wallops Island. Preliminary inspection of the telemetry records indicated all instruments functioned satisfactorily A preliminary report of the ion mass spectrometer data has already been published.[9] Analysis of the pressure gage data is not yet complete.

The mass spectrometer for measuring the neutral particle composition of the atmosphere differed from those used during the IGY,[3] in that a 7–5 cycle three-stage analyzer was used, with a resolution of about 1 part in 25. This configuration was selected instead of the four-stage IGY version to obtain greater sensitivity as well as linear operation at higher pressures. The electronic units were substantially the same as used during the IGY, and included: an emission regulator for the filament; an RF oscillator, from which the bias and stopping potentials for the analyzer were derived; a motor driven potentiometer for the voltage sweep; a 100 percent negative feedback dc amplifier to measure the ion current; and suitable regulated power supplies. The mass range of 8 to 48 AMU was swept once each 1.4

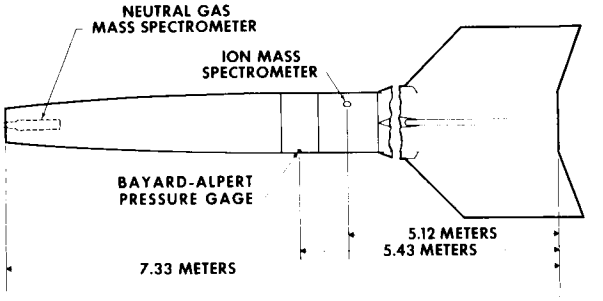


FIGURE 1.—The locations of the three principal sensors on NASA 4.14. These positions were chosen to obtain maximum isolation of the experiments from each other. The sun and earth sensors for aspect determination are on the hidden side.

* This work is described in the Consolidated Final Report on Subcontract No. 1 under Government Contract No. W 33-038 ac-14050, "Mass Spectrometer" by J. Robert Downing, Cook Research Laboratories, Chicago, Illinois, January 13, 1950. The Government contract was to the University of Michigan, College of Engineering, Department of Engineering Research, Ann Arbor, Michigan.

† Note added in proof: First successful firing, May 18-1962, preliminary data: Schaefer, Edward J., "The Dissociation of Oxygen Measured by a Rocket-Borne Mass Spectrometer" *J. Geophys. Res.* **68**(4) 1175–6, Feb. 15, 1963.

seconds. The ion source used 45-volt electrons at a nominal current of 10 ma. Other operating potentials were: RF potential, 7.7 vrms at 4.2 Mc; bias, 15.4 v; and stopping, 44.8 v.

Laboratory calibrations of the mass spectrometer included the admission of dry air through a molecular-flow type variable leak. Four calibration runs on different days with various settings of operating potentials were plotted as Ar^+/N_2^+ versus pressure over the pressure range from 10^{-6} to almost 10^{-3} mm Hg: the points had a scatter of ± 15 percent about the curve drawn in figure 2. Also plotted in figure 2 is the ratio of the argon and molecular nitrogen currents as a function of the total pressure in the mass spectrometer. This ratio was found to be independent of stopping potential and of electron emission over the range of these potentials expected in flight. The ratio was independent of pressure below about 5×10^{-5} mm Hg, but increased slowly at higher pressures as a consequence of the relatively greater probability for collisions of N_2^+ with neutral particles. The total ion current, which is the sum of the individual ion currents at the various mass

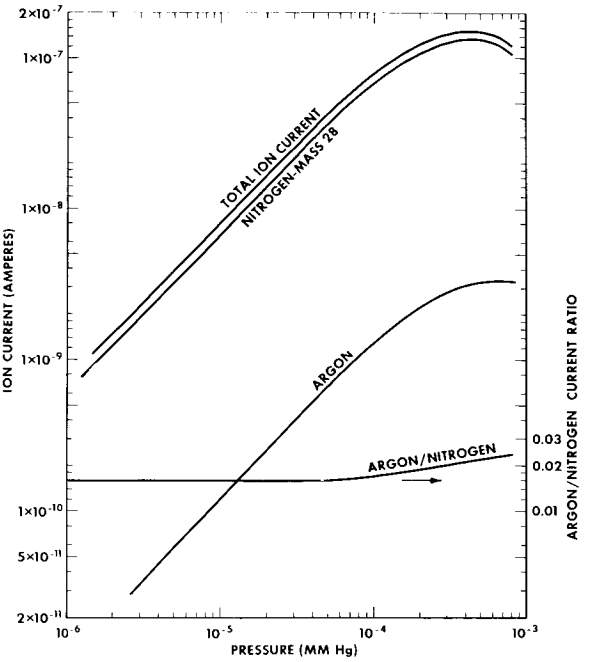


FIGURE 2.—Calibration data for the mass spectrometer flown in NASA 4.14. The curves have been adjusted so that the sensitivity corresponds to the operating conditions used in flight.

TABLE 1.—*Summary of Trajectory and Aspect Data for NASA 4.14*

Altitude (km above sea level)	Rocket total velocity (km/sec)	Ascent		Descent	
		Time (sec)	Angle of attack* (deg)	Time (sec)	Angle of attack* (deg)
100	1.545	87.5	6.3	420.4	160.8
110	1.484	94.0	9.6	414.0	159.9
120	1.418	101.0	11.7	407.0	160.9
130	1.350	108.2	12.1	399.7	163.8
140	1.280	115.8	10.7	392.0	166.9
150	1.202	124.4	7.0	384.2	168.0
160	1.123	132.8	3.1	375.2	164.9
170	1.036	142.0	5.3	365.7	159.9
180	0.942	152.5	10.3	355.7	157.2
190	0.838	163.5	12.3	344.2	159.0
200	0.722	176.8	8.8	331.3	163.1
210	0.580	192.5	6.5	315.6	158.2
220	0.393	214.1	20.5	294.1	145.3
225	0.251	231.6	26.4	276.4	139.3
227.3	0.145	253.75	82.0	253.75	82.0

* The angle of attack is the angle between the total velocity vector of the rocket and the longitudinal axis of the spectrometer.

numbers, was used as a monitor of the total pressure within the tube. The pressure calibration was obtained by means of a VG-1A ionization pressure gage located adjacent to the mass spectrometer on the vacuum system. Total spectrometer ion current was linear with respect to pressure up to 4×10^{-5} mm Hg and had its maximum value at 4.3×10^{-4} mm Hg.

After calibration of the spectrometer was completed the tube was baked in an evacuated

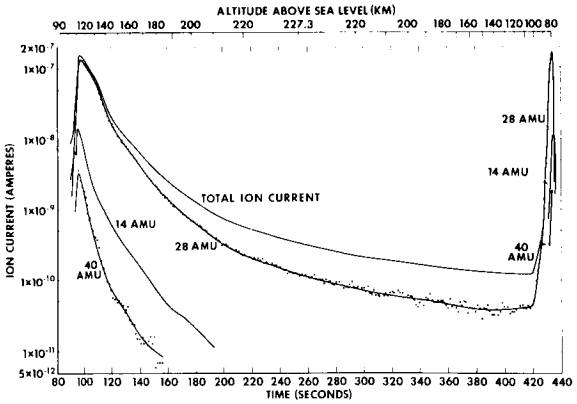


FIGURE 3.—The major atmospheric gases: molecular nitrogen at 28 AMU and argon at 40 AMU; and the total ion current which is a measure of the total pressure within the mass spectrometer.

condition and sealed. A getter, confined to the removeable tip, was flashed. The tube was then sealed with O-ring gaskets to a bulkhead at the forward end of of the rocket, about 19 inches back from the nose tip. After the top 19 inches of the rocket and the tip of the spectrometer were removed during flight at an altitude of 100 km, the opening of the spectrometer extended slightly above the bulkhead. The opening did not "see" any part of the surface of the vehicle.

DATA

The basic data concerning the flight of the rocket are given in table 1. The angle of attack is the angle between the total velocity vector of the rocket and the longitudinal axis of the spectrometer. Up to 172.4 seconds the position and velocity data were obtained from tracking radars, and the subsequent positions and velocities were computed

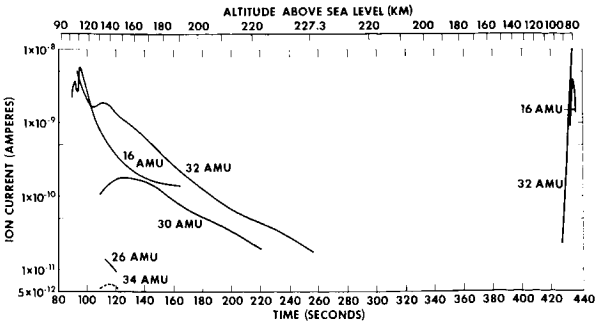


FIGURE 4.—Molecular and atomic oxygen, unidentified peaks of 26 and 34 AMU, and the nitric oxide ion produced within the mass spectrometer by an ion-molecule reaction.

for free-fall conditions. The peak altitude is known with an error of ±0.5 km. The direction of the rocket axis was determined from two rocket-mounted magnetometers and a sun-earth sensor with an accuracy of ±3 degrees. After thrust termination, the rocket's horizontal velocity was constant at 0.145 km/sec in a direction 152 degrees east of north, until the vehicle re-entered at about 85 km. The vehicle spun about its longitudinal axis at a rate of 2.69 rps and precessed in a cone of 4.6 degrees half angle; the cone axis was tipped 10.8 degrees from zenith in a direction 15.5 degrees west of north.

The basic data obtained from the spectrometer are given in figures 3, 4, 5, and 6. The output from the electrometer was telemetered on three channels each of which covered about one order of magnitude.

Figure 3 shows the spectrometer collector current due to nitrogen at 28 and 14 AMU, argon at 40 AMU, and the total ion current for these and the remaining species sensed by the spectrometer. The total does not include the

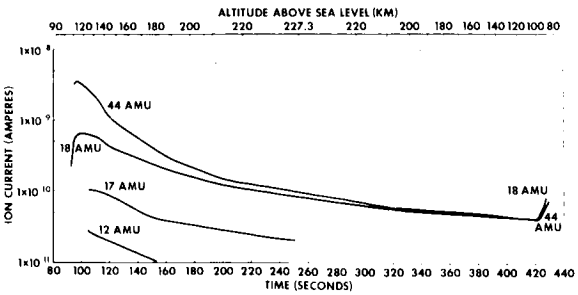


FIGURE 5.—Rocket gas: carbon dioxide and water vapor.

current due to atmospheric ions. In figure 3 the actual data points for molecular nitrogen and for argon are shown to indicate the scatter present in the various current and altitude ranges; the scatter is similar for the other constituents.

In figure 4 are plotted the ion currents for oxygen at 32 and 16 AMU, nitric oxide at 30 AMU, and some very small unidentified peaks at 26 and 34 AMU. In figure 5 are plotted carbon dioxide appearing as currents at 44 and 12 AMU, and water vapor at 18 and 17 AMU.

The ambient atmospheric ions also enter the analyzer. Because the rocket acquires a negative potential of a few volts with respect to the ionosphere, the ions and their harmonics appear and both are displaced on the mass scale. Thus, the atomic oxygen ion appears at 15.3 AMU, the molecular oxygen ion at 31.2 AMU, and the nitric oxide ion at 29.4 AMU. The lower harmonics appear in the spectra at the following positions: 32^+ at 23.2 AMU, 30^+ at 21.7 AMU, 28^+ at 20.2 AMU, and 16^+ at 11.4 AMU. The fundamental peak for ambient 28^+ is obscured by the very large peak due to neutral nitrogen. The variations of these peaks with time and altitude are shown in figure 6.

ANALYSIS OF DATA

Certain constituents of the atmosphere—nitrogen, argon, and the other noble gases—exist in fixed ratios at the earth's surface. Winds and turbulence preserve these ratios up to a high altitude. With diminishing density the speed of molecular diffusion increases, and the density profile of each permanent gas above some altitude is controlled principally by temperature and gravity. This results in a decreasing ratio of the concentration of heavy gases to that of lighter gases above the altitude at which mixing is no longer effective.

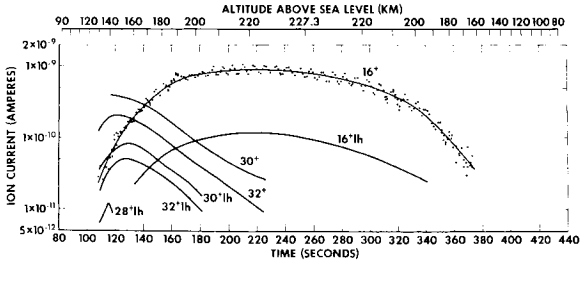


FIGURE 6.—Ambient ions and their harmonics. The fundamental peak of the 28^+ ion was hidden by the ions formed from the neutral nitrogen.

The effect of diffusive separation is noted in the mass spectra that were obtained by considering the data for argon and molecular nitrogen. The following relation between the pressure within the spectrometer P_g and the ambient density of the constituent ρ_a is used:[10]

$$\rho_a = \frac{P_g}{F(s)\,R\,(T_a T_g)^{1/2}},$$

where R=the gas constant k/m, k=Boltzmann's constant; m=mass of each gas particle; T_a=ambient temperature of the atmosphere; T_g=temperature of the gas within the spectrometer; and $F(s)=s\sqrt{\pi}[1+\mathrm{erf}(s)]+\exp(-s^2)$, in which $s=V/V_a$, V is the velocity vector of the vehicle along the spectrometer tube axis, and V_a is the most probable speed of the atmospheric constituent considered. For $s>1$, the velocity function $F(s)$ is approximately equal to $2s\sqrt{\pi}$. With this and using the expression for the most probable velocity, $V_a=(2RT_a)^{1/2}$, the ambient density is given by

$$\rho_a = \frac{P_g}{V V_g \sqrt{\pi}} = \frac{P_g}{V\left(2\pi k \frac{T_g}{m}\right)^{1/2}},$$

where V_g is the most probable velocity of the particles inside the spectrometer tube.

The foregoing discussion is valid for a chamber open to the atmosphere through an orifice. When there is a tube between the chamber and the atmosphere, as is true in this experiment, P_g must be multiplied by a function C which depends on the angle of attack, the geometry of the opening, the total velocity of the vehicle, and the accommodation coefficient of the impinging particles on the surface of the tube. For the geometry used in this experiment, C has a lower limit of $\frac{1}{2}$ when the angle of attack is 0 and a value of 1 when the angle of attack is 90 degrees. The value of C may be greater than 1 for intermediate angles of attack. It is assumed that C is the same for argon and nitrogen and will cancel when the ratio of

their densities is taken; thus

$$\left(\frac{\rho_{Ar}}{\rho_{N_2}}\right)_{flight} = \frac{P_{Ar}}{P_{N_2}}\left(\frac{28}{40}\right)^{1/2},$$

where ρ is the density of the atmosphere, P is the pressure within the spectrometer, and the subscripts Ar and N_2 refer to argon and molecular nitrogen, respectively.

Of principal interest is the comparison of the values of this ratio ρ_{Ar}/ρ_{N2} in the upper atmosphere to that for ground level air. In terms of ion current, I_{Ar} at 40 AMU and I_{N2} at 28 AMU, the ratio of the pressures in the mass spectrometer during calibration is given by

$$\frac{P_{Ar}}{P_{N_2}} = \left(\frac{\rho_{Ar}}{\rho_{N_2}}\right)_{ground} = A\left(\frac{I_{Ar}}{I_{N_2}}\right)_{ground},$$

where the subscript ground indicates the calibration with ground level air, and where A is the ratio of the sensitivity of the mass spectrometer for argon to its sensitivity for nitrogen. The variation of A with the total pressure within the mass spectrometer can be noted in figure 2 where the argon/nitrogen curve represents $P_{Ar}/P_{N2}A$. For this type of mass spectrometer A has been found to be independent of the relative abundances of nitrogen and argon when $P_{Ar}/P_{N2}<0.02$. Thus A will cancel in the following, which defines the separation ratio r:

$$r = \frac{\left(\frac{\rho_{Ar}}{\rho_{N_2}}\right)_{flight}}{\left(\frac{\rho_{Ar}}{\rho_{N_2}}\right)_{ground}} = \frac{\left(\frac{28}{40}\right)^{1/2}\left(\frac{I_{Ar}}{I_{N_2}}\right)_{flight}}{\left(\frac{I_{Ar}}{I_{N_2}}\right)_{ground}},$$

where the current ratios for flight data and ground data are taken at the same total pressure in the mass spectrometer. The total pressure in the mass spectrometer during flight was measured by the total ion current observed and the calibration curve of figure 2.

The data used to calculate the separation ratio are given in figures 2 and 3. During the period of interest, from 96 to 155 seconds, the emission current from the filament in the ion source ranged from 10.0 to 10.8 ma; however, all the plotted data have been normalized to a current of 10.0 ma. During the first few seconds of operation of the spectrometer tube in flight, that is, from 89 to 95 seconds, a glow discharge apparently occurred in the ion source, with the result that the values of ion currents were not proportional to emission current and resolution of the spectrometer was impaired during that interval of time. The data from these spectra were not used in the computations.

The result, the separation ratio, is shown in figure 7. The points are derived from the values of the smooth curves in figure 3, read at the altitudes in question. The error in this curve (figure 7) is judged to be no more than ±20 percent from systematic and reading errors. The reading errors rise rapidly with altitude above 150 km; the extension of the separation ratio is shown in figure 9 only to indicate the trend.

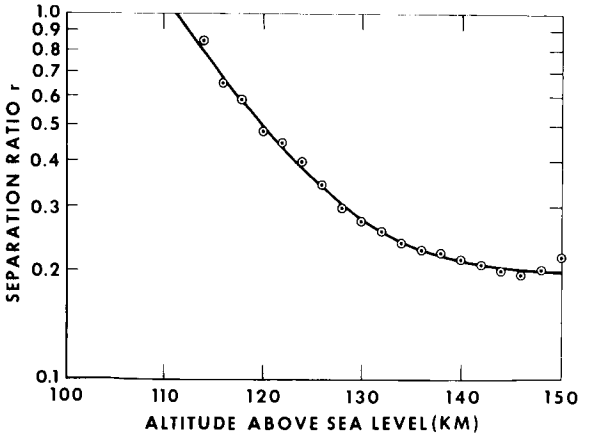


FIGURE 7.—The separation ratio r is the value of the ratio of the argon and molecular nitrogen densities measured in flight compared to the value for the same ratio measured for ground level air. This curve is for NASA 4.14, November 15, 1960, 1141 EST.

DISCUSSION

Ambient Ions

Atmospheric ions also enter the spectrometer and appear in the spectra, displaced about one half a mass unit from ions of the same mass produced in the ion source. The displacement is due to the initial energy the ions acquire from the negative potential that the vehicle has with respect to the ionosphere. The data for these ions

are shown in figure 6. The experimental data for atomic oxygen ions are given to show the cyclic variation of 15 to 20 percent with time. This correlates well with the phase of the roll angle. The reason for this pattern is not clear, for the experiment was intended to be symmetrical about the roll axis, and no similar variation was noted for the neutral particles. One possible explanation is that a trace of foreign material on one side of the interior of the spectrometer tube entrance yielded photoelectrons copiously, reducing the number of ions which successfully reached the analyzer. Such a phenomenon was noted in the data from the ion spectrometer flown in Viking 10 (May 7, 1954), in which the ion current was correlated with the angle of the sun.[11] Another possible explanation assumes a nonsymmetrical distribution of magnetically permeable material near the entrance such that the resultant field, as modified by the material's interaction with the earth's magnetic field, depended on the phase of the roll angle. Hence, the mean free path and density of the electrons in the entrance tube may have changed in response to the varying magnetic field altering the probability of ambient ions reaching the analyzer.

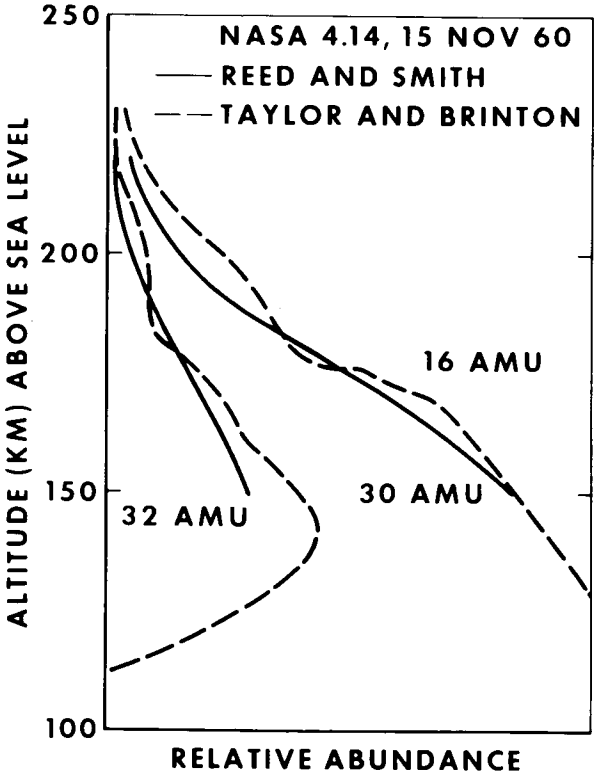


FIGURE 8.—Ion composition measured by NASA 4.14. The relative width of the horizontal space allotted to each species indicates its relative abundance. Not included is the trace of the 28 AMU ion noted between 132 and 142 km.

The relative altitude distribution of the ambient ions measured by the neutral gas spectrometer is shown in figure 8. These data support the measurement of ion composition by the ion spectrometer on this vehicle as reported earlier by Taylor and Brinton,[9] and as indicated in figure 8.

Other Gases

Little useful information about the upper atmosphere was obtained from the remaining ion current peaks. Mass number 14 (N^+) appears (figure 3) as the result of dissociative ionization of molecular nitrogen within the spectrometer's ion source as well as from ionization of atomic nitrogen. All the mass 14 ion current in these spectra is attributed to molecular nitrogen. Any atmospheric atomic nitrogen either disappeared on the surfaces of the spectrometer, or its abundance is less than 4 percent of that of molecular nitrogen.

No completely plausible explanation has been found for the shape of the altitude profiles of mass numbers of 16(0^+) and 32(0_2^+) in figure 4. It may be that immediately after the spectrometer was opened to the atmosphere both atomic and molecular oxygen were adsorbed on the surface of the spectrometer. After 120 seconds, the principal source of molecular oxygen within the spectrometer may be the gas released from these surfaces, the supply to the surface being constantly replenished by the oxygen atoms and molecules from the atmosphere. It does not seem possible to deduce from these data the relative abundance of atomic and molecular oxygen either with respect to each other or to the total atmosphere. The nitric oxide peak (30 AMU) comes from charge exchange phenomena within the ion source of the mass spectrometer.[12]

Carbon dioxide and water vapor, represented by 44 (CO_2^+) and 18 (H_2O^+) in figure 5, have been consistently observed in the spectra from the various rocket flights of the mass spectrometer, and are presumed to come from the vehicle. There is also in the spectra the secondary peak due to

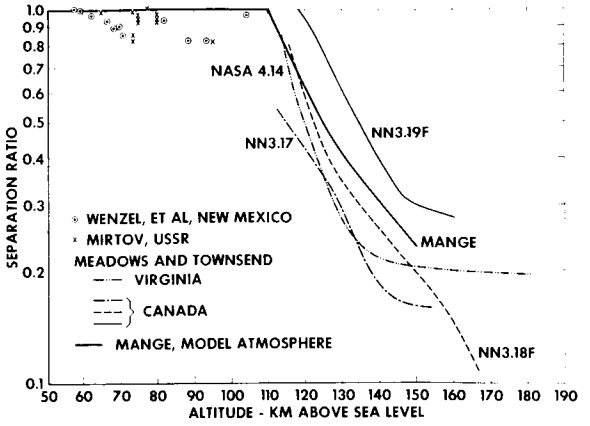


FIGURE 9.—Summary of data concerning the relative abundance of argon and nitrogen at various altitudes.

water vapor (OH^+ at 17 AMU) with an ampltiude of about 0.3 that of the 18 AMU ion current peak, and a small peak at mass 12, probably due to C^+ from the carbon dioxide. Other small peaks occurred at mass numbers 34 and 26 (figure 4). Their chemical composition and origin have not yet been explained.

Diffusive Separation

A summary of data concerning the distribution of argon and nitrogen is shown in figure 9. All the data below 110 km were obtained from the laboratory analysis of gas samples collected by rocket borne sample bottles flown at temperate latitudes. The first American samples were obtained from bottles flown on Aerobee rockets by Jones et al., of the Department of Aeronautical Engineering at the University of Michigan. The data shown in figure 9 are from samples obtained from four Aerobee rockets and a V-2 rocket flown between 1950 and 1956 at White Sands Proving Ground in New Mexico. The samples, collected in steel bottles, were divided and analyzed by the group at the University of Michigan and by Professor Paneth of Mainz by separation of the various constituents. A comprehensive series of articles describing these experiments appears in the literature;[13–16] a summary article of the program was written by Wenzel, Loh, et al.[17]

Scientists of the USSR collected a number of air samples on rocket flights from 1951 through 1956. Their techniques of collection and analysis were quite different from those employed in the United States. The containers were of glass and the analyses were carried out by studying the emission lines of the gas when subjected to a high frequency discharge. Their results, as presented in the summary article by Mirtov,[18] are also plotted in figure 9.

Of the flights of the radio frequency mass spectrometer prior to the IGY program only one, Aerobee NRL-13, yielded acceptable spectra. There was no evidence of diffusive separation of argon and molecular nitrogen in these data, at a maximum altitude of 137 km.[2] However, the nitrogen pressure in the spectrometer was about a factor of 10 greater than that which would be predicted from the various model atmospheres. It is now believed by the authors that the data obtained constituted a better measure of the composition of the air escaping from the interior of the rocket's nose cone than of the ambient atmosphere.

The information concerning the distribution of argon and nitrogen obtained with radio frequency mass spectrometers flown by the USSR scientists has been summarized by Pokhunkov.[19] He found that both in the day and night atmospheres there is gravitational separation of argon and nitrogen. At night (midnight, September 9, 1960) this separation begins in the 105 to 110 km altitude region. The definition of this level in the day atmosphere was difficult: in the first two flights (morning, July 14 and 22, 1959) there was a limited number of data points near 100 km; in the last flight (late afternoon, November 15, 1961) the experiment was placed in operation and began to record spectra at 130 km. However, the data observed were in agreement for all three experiments and show appreciable separation of argon and nitrogen.

The data above 110 km shown in figure 9 include the results from the mass spectrometers mounted on the pre-IGY rocket NN3.17 and the two IGY rockets NN3.18F and NN3.19F flown at Fort Churchill, Canada.

The pre-IGY flight occurred at 2321 CST on November 20, 1956. Although auroral activity was observed during the evening and ground magnetograms indicated a disturbed magnetic condition,

an overcast sky prevented visual auroral observations at the time of firing. Though there may have been some auroral activity then, there was no evidence of a major display in progress. Care should be taken in extrapolating the curve of NN3.17 toward lower altitudes: Since the temperatures in this region are low (200° to 250°K), it is expected that the separation ratio would attain a value of 1 somewhat above 100 km. The second firing, NN3.18F, was at 2002 CST on February 21, 1958. An auroral display observed visually was too weak and diffuse for ground camera recording; it is not known whether the vehicle entered the aurora.

It is believed that these two flights (NN3.17 and NN3.18F) can be considered typical of the arctic nighttime atmosphere.

The third flight, NN3.19F, was carried out at 1207 CST on March 22, 1958, during a polar blackout. An additional influx of energy, in the form of 10 to 100 Mev particles of x rays or both, is associated with polar blackout. The absorption of this energy results in heating and turbulence, possibly changing both the extent of mixing and the temperature profile of the atmosphere.

The last flight, NASA 4.14, at 1141 EST, was made at a temperate latitude; ionograms indicated that the ionosphere was becoming quiet after a period of moderate disturbance. However, the Fredericksburg Magnetic Observatory at Corbin, Virginia, recorded a moderately severe magnetic storm from 0804 to 1221 EST[20] and the sunspot number was high, 133.[21]

Both Nicolet[22] and Mange[23] have considered the distribution of argon as modified by the process of diffusion. Mange computed the altitude dependence of various atmospheric constituent concentrations in diffusive equilibrium above an arbitrary altitude of 110 km, with a temperature of 318°K at 120 km, and a temperature increase with height of 6 deg/km. The curve labeled Mange in figure 9, shown for comparison with the experimental data, is obtained from the computation by Mange presented in figure 1 of his paper.

CONCLUSION

In conclusion, it may be said that the level at which diffusive separation of argon and molecular nitrogen becomes effective has been measured and found to vary from about 100 to 119 km, with typical values between 110 and 112 km.

ACKNOWLEDGMENTS

The authors express their appreciation to the people who made the NASA 4.14 flight possible, in particular to Dr. John W. Townsend, Jr., who laid down the basic design of the experiment some years ago and who made some very helpful suggestions in the analysis of these data; to Mr. Richard G. Holt and Mr. Alfred J. Heine, Jr. who assisted in all phases of the construction and preparation of the mass spectrometer instrumentation; to Mrs. Nancy Biess who took special care in the reading of telemetry records and subsequent computations; and to Mr. Harry Taylor who so ably stepped in, upon short notice, as Scientific Officer during the final preparations and firing of the rocket.

REFERENCES

1. Smith, C. P., Jr., "Upper Atmosphere Research Report No. XXI—Summary of Upper Atmosphere Rocket Research Firings," U. S. Naval Research Lab. Rept. No. 4276, February 1954; Supplement by E. C. Pressly, February 1958.
2. Townsend, J. W., Jr., Meadows, E. B., and Pressly, E. C., "A Mass Spectrometric Study of the Upper Atmosphere," in: *Rocket Exploration of the Upper Atmosphere*, ed. by R. L. F. Boyd and M. J. Seaton, London: Pergamon Press, 1954, pp. 169–188.
3. Meadows, E. B., and Townsend, J. W., Jr., "IGY Rocket Measurements of Arctic Atmospheric Composition above 100 Km," in: *Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January 1960*, ed. by H. K. Bijl, Amsterdam: North-Holland Publ. Co., 1960, pp. 175–198.
4. Pokhunkov, A. A., "Mass-Spectrometric Investigations of the Structural Parameters of the Earth's Atmosphere at Altitudes of from 100 to 210 Kilometers," *Iskusstvennye Sputniki Zemli* **7**: 89–100, 1961 (In Russian); Translation in *Artificial Earth Satellites*, **7** and **8**: 88–100, Plenum Press, New York, June 1962.
5. Istomin, V. G., and Pokhunkov, A. A., "Mass Spectrometer Measurements of the Atmospheric Composition in the USSR," in: *Space Research III: Proc. 3rd Internat. Space Sci. Sympos., Washington, May 1962*, ed. by W. Priester, Amsterdam: North-Holland Publ. Co., 1963 (in Press).
6. O'Day, M. D., "Upper Air Research by Use of Rockets in the U. S. Air Force," in: *Rocket Exploration of the Upper Atmosphere*, ed. by R. L. F. Boyd and M. J. Seaton, London: Pergamon Press, 1954, pp. 1–10.

7. JOHNSON, C. Y., "Aeronomic Parameters from Mass Spectrometry," *Annales de Geophysique* **17**(1): 100–108, 1961.
8. LAGOW, H. E., HOROWITZ, R., and AINSWORTH, J., "Results of IGY Atmospheric Density Measurements above Fort Churchill," in: *Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January 1960*, ed. by H. K. Bijl, Amsterdam: North-Holland Publ. Co., 1960, pp. 164–174.
9. TAYLOR, H. A., JR., and BRINTON, H. C., "Atmospheric Ion Composition Measured above Wallops Island, Virginia," *J. Geophys. Res.* **66**(8): 2587–2588, August 1961.
10. HOROWITZ, R., and LAGOW, H. E., "Upper Air Pressure and Density Measurements from 90 to 220 Kilometers with the Viking 7 Rocket," *J. Geophys. Res.* **62**(1): 57–78, March 1957.
11. JOHNSON, C. Y., and MEADOWS, E. B., "First Investigation of Ambient Positive-Ion Composition to 219 Km by Rocket-Borne Spectrometer," *J. Geophys. Res.* **60**(2): 193–203, June 1955.
12. POTTER, R. F., "Ion-Atom Interchange in Air," *J. Chem. Phys.* **23**(12): 2462–2463, December 1955.
13. JONES, L. M., ET AL., "The Measurement of Diffusive Separation in the Upper Atmosphere," in: *Rocket Exploration of the Upper Atmosphere*, ed. by R. L. F. Boyd and M. J. Seaton, London: Pergamon Press, 1954, pp. 143–156.
14. PANETH, F. A., "The Chemical Analysis of Atmospheric Air," in: *Rocket Exploration of the Upper Atmosphere*, ed. by R. L. F. Boyd and M. J. Seaton, London: Pergamon Press, 1954, pp. 157–158.
15. REASBECK, P., and WIBORG, B. S., "Chemical Analysis of Upper Atmosphere Air Samples from 50 km to 93 km Height," in: *Rocket Exploration of the Upper Atmosphere*, ed. by R. L. F. Boyd and M. J. Seaton, London: Pergamon Press, 1954, pp. 158–161.
16. MARTIN, G. R., "The Composition of the Atmosphere Above 60 km," in: *Rocket Exploration of the Upper Atmosphere*, ed. by R. L. F. Boyd and M. J. Seaton, London: Pergamon Press, 1954, pp. 161–168.
17. WENZEL, E. A., LOH, L. T., ET AL., "The Measurement of Diffusive Separation in the Upper Atmosphere," in: *Annals of the International Geophysical Year*, London: Pergamon Press, 1960, Vol. 12, Part 1, pp. 407–421.
18. MIRTOV, B. A., "Rocket Investigations of Atmospheric Composition at High Altitudes," *Uspekhi Fizicheskikh Nauk* **63**(1b): 181–196, September 1957 (In Russian); Translation in *The Russian Literature of Satellites*, New York: International Physical Index, 1958, Part II, pp. 67–85.
19. POKHUNKOV, A. A., "Gravitational Separation, Composition and Structural Parameters of the Atmosphere at Altitudes Above 100 Km," in: *Space Research III: Proc. 3rd Internat. Space Sci. Sympos., Washington, May 1962*, ed. by W. Priester, Amsterdam: North-Holland Publ. Co., 1963 (In Press).
20. LINCOLN, J. V., "Geomagnetic and Solar Data," *J. Geophys. Res.* **66**(3): 979–981, March 1961.
21. LINCOLN, J. V., "Geomagnetic and Solar Data," *J. Geophys. Res.* **66**(4): 1279–1285, April 1961.
22. NICOLET, M., "The Properties and Constitution of the Upper Atmosphere," in: *Physics of the Upper Atmosphere*, ed. by J. A. Ratcliffe, New York: Academic Press, 1960, pp. 17–71.
23. MANGE, P., "Diffusion in the Thermosphere," *Annals de Geophysique* **17**(3): 277–291, 1961.

RESEARCH WITH TIROS RADIATION MEASUREMENTS

WILLIAM NORDBERG

Goddard Space Flight Center

The advantages of viewing the earth and its atmosphere from an orbiting meteorological satellite have been thoroughly expounded by now and are well known. The Tiros series has demonstrated the usefulness to the meteorologist of observations of large-scale cloud cover and similar features with television cameras. TV observations, while providing a maximum resolution of the areas viewed, nevertheless have two distinct disadvantages: Their response is limited to a rather small portion of the visible spectrum (0.5–0.7 microns) and the absolute intensities of the radiation received by the camera cannot be measured. The greatest asset of TV observations therefore lies in the ability of the cameras to map meteorological features with relatively high resolution.

Much more can be learned, however, about the physical structure of the atmosphere and its meteorological implications by extending observations to other portions of the spectrum, particularly to the infrared, and by obtaining a more precise measurement of intensities of radiation emanating from various regions of the globe. This makes it highly desirable to supplement TV photographs with radiometric observations.

A radiometer generally consists of a photodetector (the nature of which will largely depend on the desired spectral response), an optical filter and, possibly, light gathering lenses or mirrors. It has to be well calibrated in terms of its irradiance, but a variety of spectral regions may be chosen by selecting the proper components, such as transmission filters, reflective coating, etc. All measurements discussed here resulted from rather simple instruments which reflect the state of satellite radiometry of several years ago, when the idea of this type of experiment was first realized.

Today, one might conceive of considerably more advanced, high-resolution radiometric sensors, some operating in the extremes of the electromagnetic spectrum such as microwave radiometers, but their concepts are described elsewhere in this issue [This work, Vol. II (Space Technology), p. 563]. To some extent, the design of such future devices will undoubtedly be based on the results reported here.

Radiometric experiments have been generally performed by meteorological satellites for one or a combination of the following three reasons:

1. To map the distribution of cloud patterns both day and night and determine heights of cloud tops. The simplest device to accomplish this is a radiometer operating in a narrow portion of the infrared spectrum where the transmission through clear atmosphere is a maximum (that is, an atmospheric "window"). One of the most effective windows for this purpose lies in the 10–11 micron region. The radiation received at the instrument is assumed to be a known function of the effective temperature only of the emitting surface (ground or cloud top). Since temperature distribution with height in the troposphere is assumed to be known from balloon or climatological data, radiation measurements can not only detect the presence or absence of clouds, but also determine the heights of the cloud tops. Two requirements must be fulfilled by the radiometer: The field of view of the instrument must be reasonably small to provide sufficiently high spectral resolution (0.5 and 3.0 deg are used in present high-

and medium-resolution instruments), and the field of view must scan the earth continuously to give adequate coverage.

2. To determine the structure of the atmosphere. Here again, the radiometer must operate within a narrow spectral region near the center of a strong absorption band of one of the atmospheric constituents (6.7 microns for H_2O or 15 microns for CO_2). The radiation received is a function of both the temperature and the concentration of the constituent with height. By measuring radiation in the absorption band and comparing it with measurements in a window channel, the temperature structure of the lower atmosphere can be derived if the concentration of the absorbing gas is known (as in the case of CO_2), or the total content of the absorber in the viewed column can be determined if the temperature structure is known. Total amounts of water vapor in the column may be determined by this last method. Because of the high spectral resolution necessary, it has not been possible so far to use the former method. Such an instrument is now under development.

3. To determine the balance between solar energy absorbed and thermal energy emitted by the earth and the atmosphere. This balance is a major factor in determining the energy stored in the atmosphere and available to be transformed into dynamic processes. Measurements for this purpose may be performed by two extremely simple sensors. One must operate in the solar spectrum (0.2–4 microns) to detect reflected solar energy from the earth, and the other in the infrared (5–30 microns) to receive emitted thermal radiation. Spatial resolution in this case is not of the essence, and the field of view may be one order of magnitude larger than for a medium-resolution scanning radiometer.

Each of these three areas holds great importance for meteorology. A complete mapping of cloud distribution and heights is of immense importance to the synoptic meteorologist: a knowledge of global height distributions of temperature and water vapor could eventually find its way into schemes for numerical weather forecasting, and the measurement of radiative energy balance might provide a powerful tool to explore the formation of storm systems.

Of the three areas, the first has been most extensively satisfied by measurements from all meteorological satellites in the past[1–3] and will also be strongly pursued in the near-future with Nimbus experiments.[4] Measurements from Tiros[5–8] and Explorer VII have been highly encouraging with respect to measurement of total energy fluxes, and some results were obtained from Tiros III and IV water vapor in the atmosphere.[9]

In addition, these experiments have produced a wealth of secondary results, such as background information for the better use of horizon scanners,[10] the durability of optical components in the space environments,[11] the development of useful infrared calibration techniques in the laboratory,[12] and the development and use of a magnetic attitude-control device.[13] The experiments and their basic results are reviewed below.

EXPERIMENTS TO DATE

The first meteorological satellite, Vanguard II, carried as its only payload two PbS photocells with appropriate optical systems, electronics, tape recorder, and telemetering. The photocells were to scan the earth to map cloud cover. The scanning motion was to be produced by the regular spin of the satellite.[14] Unfortunately, the satellite tumbled and no systematic maps could be generated. However, the experiment was successful inasmuch as it demonstrated the capability of this technique and paved the way for the subsequently successful Tiros radiation experiments.

The sensors flown in Tiros II, III, and IV have been fully described elsewhere.[15] Each satellite contains two instruments. On is a nonscanning radiometer with broad response in both the visible and infrared regions and a rather low spatial resolution. Its field of view is approximately 55 deg. The other is a scanning radiometer which scans as the satellite spins. It is of medium spatial resolution, has a 5-deg field of view, and responds to radiation in five different spectral regions determined by optical filters. Three of these regions are in the infrared between 5.9 and 6.7 microns, 8 and 12 microns, and 8 and 30 microns. The other two lie mainly in the visible portion of the spectrum between 0.50 and 0.75 microns and 0.2 and 7.0 microns. These spectral regions are purely nominal since the response of the instrument within these regions is far from uniform. There-

fore, the exact spectral response curves must be used when the energy distribution in each channel is calculated. Typical curves for Tiros III have been given elsewhere.[16]

In the 5.9 to 6.7 micron channel, a maximum of absorption due to water vapor is encountered. Energy in this channel is therefore received mainly from the highest altitudes where water vapor may be found in the atmosphere. In contrast, the optical depth is a maximum in the 7.5 to 13.5 channel, since absorption due to any of the atmospheric constituents—except ozone, which covers only a minor portion of this channel—is very small. The 7.0 to 32.0 channel covers almost 80% of the total black-body energy emitted by the earth, while the total solar energy reflected from the earth is contained in the 0.20 to 7.0 channel. The 0.50 to 0.75 channel is of interest because it covers only a narrow spectral region very near the maximum of solar-energy distribution, and is similar in its spectral response to the TV cameras carried on the same satellite.

One of the basic differences between the wide-field and five-channel instruments is that the five-channel radiometer scans the earth during all portions of the orbit, while the wide-field radiometer fully views the earth during less than $\frac{1}{5}$ of the orbit. The wide-field instrument measures the radiant emittance of a target by means of a thermistor whose resistance is a function of the absorbed energy flux. This is accomplished over the whole spectrum with two detectors: a black one equally sensitive to radiation emitted and reflected from the earth, and a white one predominantly sensitive to emitted radiation. The portion of solar energy reflected from the earth (the earth's albedo) and the apparent black-body temperature for earth can be determined by comparing the energies received by the black and white detectors. A complete description of this instrument has been given by Hanel.[17]

Very much in contrast to the wide-field radiometer, the sensors in the five-channel instrument are alternately, and in rapid succession, illuminated with two diametrically opposed fields of view, one scanning the earth and the other pointing into outer space. The sensors therefore measure the difference of the energy fluxes in the two directions. Since the flux from outer space is essentially zero, this serves as a reliable reference.

A detailed description of this instrument and its scanning mechanism is contained in.[15] Further instrumental details, such as recording, transmitting, and electrical conversion of the measurements have been described.[18]

In addition, Tiros III and IV each carried radiation sensors of the type used in Explorer VII and described below.[21]

On Explorer VII, incident radiation from the sun, reflected solar radiation from the earth, and emitted radiation from the earth are measured with simple bolometers in the form of hollow silver hemispheres. The hemispheres are thermally isolated from, but in close proximity to, specially aluminized mirrors. The image of the hemisphere which appears in the mirror makes the sensor look like a full sphere. The mirror is made large enough so that no part of the hemisphere bolometer "sees" the satellite itself. The temperatures of the thin silver hemispherical bolometers are measured by glass-coated bead thermistors, mounted so as to provide good thermal contact to the hemisphere. In addition, provision is made to measure the temperature of the mirrors.[21]

It is possible to show that a mirror-backed hemisphere by virtue of the satellite spin acts, as far as the radiation fluxes are concerned, very similar to an isolated sphere in space. Two hemispheres are coated black, which makes them respond about equally to solar and terrestrial radiation. Another hemisphere, coated white, is more sensitive to terrestrial radiation than to solar radiation. A fourth, with a gold surface, is also more sensitive to solar radiation. A black sphere, mounted on the axis of the satellite at the top, is used to determine any deterioration in the mirror surfaces by comparison with the blackened hemispheres. Finally, a small Tabor-surfaced hemisphere equipped with a shade to protect it from direct sunlight can be used to measure reflected sunlight when the axis of the satellite points to the earth's surface.

The information telemetered to the earth's surface is sensor *temperatures*. The radiation fluxes are obtained by using these temperatures in heat-balance equations.

Two scanning radiometers similar in concept to

the Tiros experiments will be flown in Nimbus. The radiometers have now reached the prototype stage. One, a high resolution radiometer with a $\frac{1}{2}$-deg field of view, will map thermal radiation in an atmospheric window between 3.5 and 4.2 microns.

The other, a medium-resolution radiometer, is a continuation of the Tiros five-channel radiation experiment. Early versions will have about the same wavelength ranges as the Tiros radiometer but will be improved. The spectral range of the 8–12 micron channel will be narrowed to a 1-micron band from approximately 10 to 11 microns, where the window is most transparent. Spectral response of other channels has also been improved.

The major advance, however, is in the check of calibration during operation in space. Both medium- and high-resolution radiometers will scan the radiometer structure, which has been converted into a black body by deep grooves and a proper coating, and which has a separate telemetering channel to monitor its temperature. Now two points on the calibration curve are available for all thermal (infrared) channels—zero (outer space) and a point close to the maximum of the dynamic range. However, some channels which are insensitive to infrared, such as the 0.25- to 4-micron channel of the medium-resolution radiometer, need other means of checking the calibration. Here the sun provides the calibration signal.

Twice each orbit, the lower surface of Nimbus is exposed to sunlight, just before and just after spacecraft enters and leaves the earth's penumbra. A metal reflector and a sapphire bead then channel sunlight into the view of the radiometer.

In contrast to Tiros, where scanning motion is provided by satellite spin, Nimbus radiometers generate the scan internally, by a rotating mirror. Scan lines are perpendicular to the velocity vector of the satellite and provide complete coverage of the earth from horizon to horizon without overlaps or gaps in the scan pattern at the subsatellite point. Use of the atmospheric window from 3.5 to 4.2 microns made the application of fast semiconductive detectors possible without going to excessive cooling requirements. In the satellite, a radiative cooling system maintains a detector temperature of approximately +80°C, sufficient for the lead selenide detector cell used in the high-resolution radiometer. The cooler is a surface of high emissivity, $2\frac{1}{2}$ by 4 cm in size. It is thermally in contact with the cell but isolated from the spacecraft, exposed to outer space but never to the sun. This is possible on Nimbus since the orientation of the spacecraft is actively controlled with respect to earth and sun.

Aside from engineering the instruments, there are two major challenges common to all radiometric experiments: Calibration of the sensors and handling the vast amounts of information continually accumulated by the satellite. These challenges are particularly pronounced in the Tiros five-channel radiometer experiment, where each channel must not only be calibrated in terms of its total irradiance but the precise spectral reponse must also be known since in one orbit the instrument feeds over one million data bits to a digital computer.

The calibration for total irradiance is accomplished with specially constructed black bodies in vacuum for the thermal channels and with precision calibrated tungsten sources for the solar-radiation channels. Calibration for spectral irradiance becomes a major difficulty in the long wave end of the 8–30 micron channel. In the visible channels, this calibration must be performed with great precision because the tungsten calibration target and the reflecting earth have vastly different temperatures (2800 and 6000°K).

Data are recorded on a magnetic tape containing the radiation intensities in digital form as a function of geographic location, viewing geometry and time. This record, called FMRT (Final Meteorological Radiation Tape), is produced by a highly complex program in an IBM 7090 computer from digitized telemetry records and a knowledge of the calibration of the sensors and the satellite's location and attitude. The latter is especially cumbersome to incorporate since none of the satellites flown so far has been earth-oriented. In utilizing the data that is, in plotting cloud cover maps, deriving energy budgets, or performing other research on radiation data, it is most advantageous to program the problem and operate directly on the FMR tape without intermediate printouts.

For those who do not have easy access to large digital computers data catalogs were published on Tiros II[19] and III[20] which contain a representative sample of data collected, and are intended to show the potential of the radiation measurements. The map (Fig. 1) is a typical example of such data display. Its significance will be discussed below.

REVIEW OF EXPERIMENTAL RESULTS

The wide-field radiometers on Explorer VII[3] have already given an indication that patterns of emitted radiation fluxes from the earth can be well correlated with cloud patterns. It was the Tiros five-channel instrument, with its much greater resolution, however, which demonstrated most convincingly its ability to map cloud cover. To date, innumerable cases have been analyzed where radiation in the 8–12 micron window channel was compared with cloud photographs and other synoptic data and the agreement was found to be excellent.

A typical example is depicted in the illustrations figure 2. The map shown in the top illustration was drawn automatically by the computer directly from the FMR tape, and is typical for the results presented in the Radiation Data Catalogs.[19,20] In the other illustration, the same map has been contoured manually for better clarity, with regions of different black-body temperature ranges shown in different shades of grey, and the path of the satellite subpoint from northwest to southeast also indicated. The region of minimum temperature (200–230°K) near 13°N, 72°W indicates a circular cloud system which, because of its very low temperature, must extend to extreme altitudes. Indeed, this system coincides with hurricane Anna, identified through cloud photographs and aircraft reconnaissance. In a series of similar radiation maps, the storm can be tracked over its entire path almost 4000 km.

Because of the small scale of plotting, the temperatures shown on the maps generally represent averages over several scan spots withon one grid element. Therefore the maps do not necessarily show the true minimum temperature measured in the storm center. In fact, careful examination of each scan spot reveals that this temperature ranged between 200 and 210°K. From a climatological temperature profile, shown in Fig. 5, one finds the height of the cloud top for the hurricane at approximately 15 km, while the cloud system just to the southeast of the storm shows minimum temperatures near 230°K and therefore a cloud top height of approximately 11 to 12 km.

Such height determinations can only be made from radiometer data and are not at all evident from cloud photographs such as the one shown at left here. Figure 3 shows a Tiros photograph over hurricane Anna taken at the time of the radiation measurements of the two illustrations on page 77. In addition, manually reduced contour lines of radiation values in the 8–12 micron channel prepared from the *analog* telemetering records are superimposed on the photograph. The analysis of these analog radiation data and the contouring of

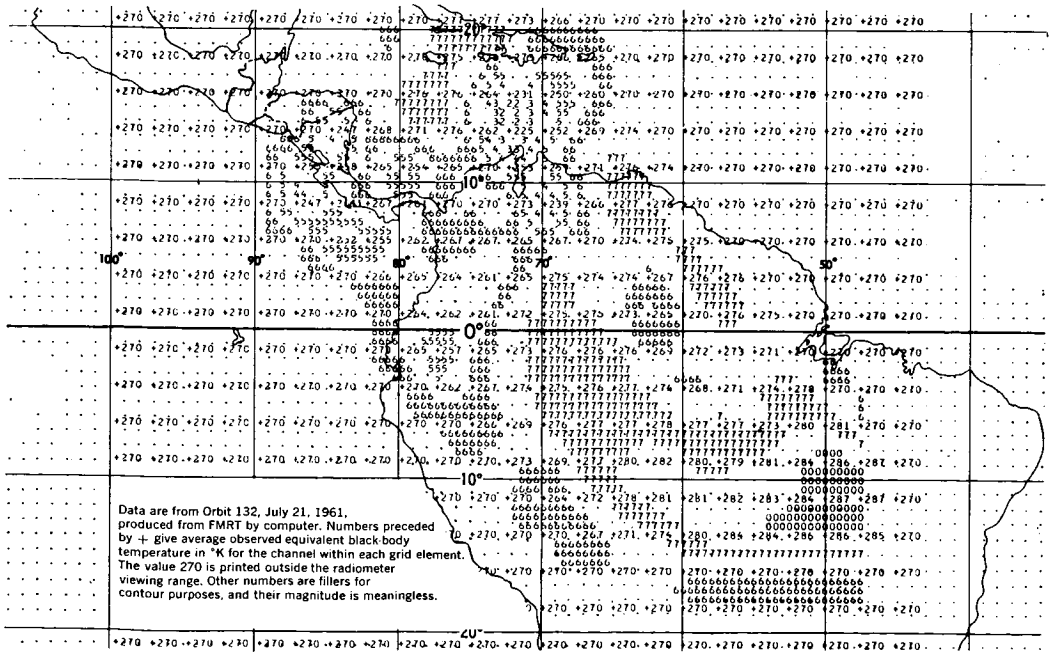


FIGURE 1.—Tiros III radiation maps, 8–12 micron channel—top, produced by computer; bottom, reduced manually.

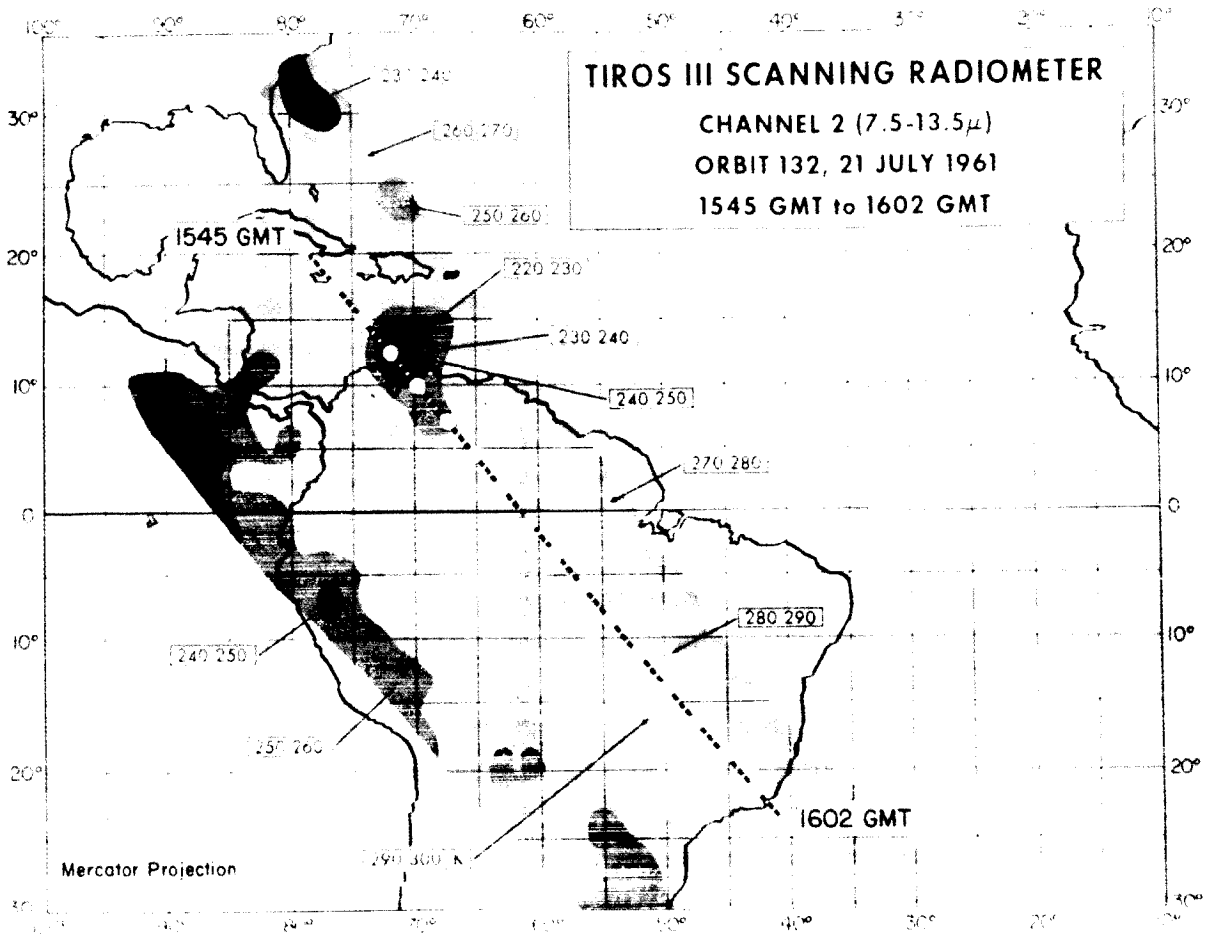


FIGURE 2.—Data are from orbit 132, with radiation values from FMRT and contours drawn manually. Numbers identifying gray shades correspond to equivalent black-body temperatures in °K. Satellite path is indicated by dashed line, with white circles marking its position every full minute.

the photograph shown adjacent was carried out by Prof. T. Fujita and his staff at the Univ. of Chicago. His kind permission to reproduce his results in this article is gratefully acknowledged.

The perfect match between radiation and cloud patterns is striking. It is this type of presentation, given on an incomparably larger scale than in the two maps, which demonstrates the full cloud-mapping ability of the radiation sensors. Unfortunately, in the illustration at the left, the radiation intensities were plotted on a relative scale only and the contour numbers shown are arbitrary and not well suited for cloud height determinations.

Additional features apparent in the map on page 1429 are the huge clear area (290 300°K) over central South America corresponding to the high-pressure system one would expect to find in this subtropical belt in the winter hemisphere, and the large cloud mass off the coast of Florida. Comparisons such as the one shown in the map on page 1431 and the ability of computers to create meaningful maps from the digitized data clearly demonstrate that radiometric observations in an infrared "window" region may well be used for synoptic analysis in future real-time meteorological satellite systems.

It must be strongly emphasized, however, that such systems are still very much in the future. The example given for the determination of cloud-top heights illustrates the principle only and any cloud height measurement with present Tiros instruments is very approximate. Many difficulties of a theoretical nature must still be overcome. For example, no true atmospheric window really

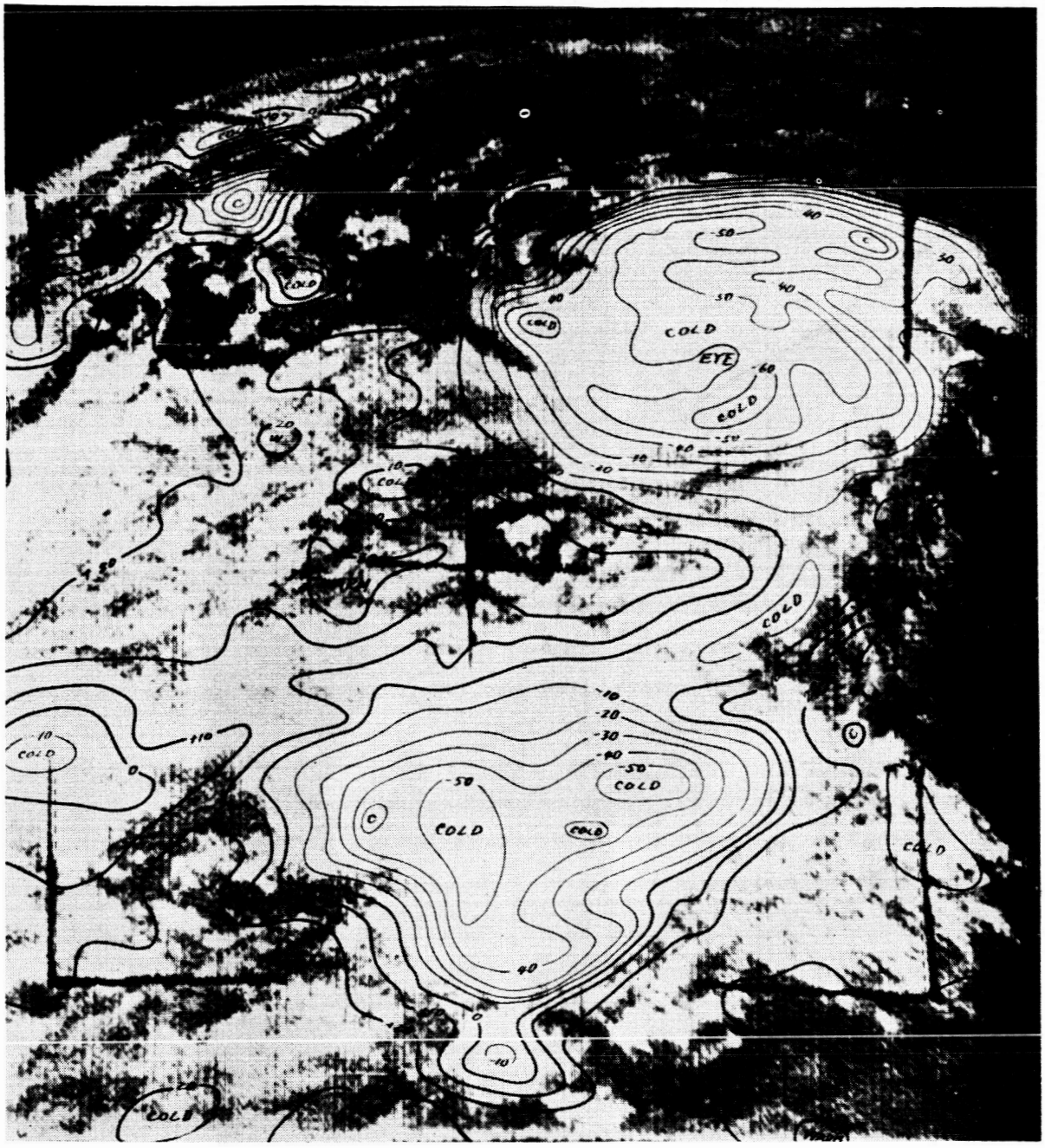


FIGURE 3.—Hurricane Anna as seen by Tiros III (orbit 132) near 13°N, 72°W. Contour lines represent relative radiation intensities from 8–12 micron channel. Radiation contours drawn from raw analog data.

exists where the transmission even in clear atmosphere is 100%. Black-body temperature measurements in the 8–12 micron channel by Tiros must therefore be corrected for absorption by water vapor, carbon dioxide, and ozone. A theory to apply such corrections was developed by Wark and Yamamoto.[2] Even after applying these corrections, investigations show that, especially when cloud tops are near the earth's surface, radiation measurements yield temperatures 5–10°K too low.

There could be several reasons to explain this:

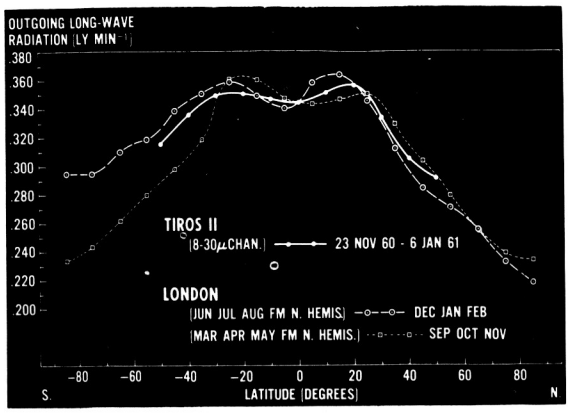


FIGURE 4.—Outgoing long-wave radiation vs. latitude, measured and predicted. Tiros II results (solid curve) are compared with theoretical curves of J. London (dashed curves).

Clouds which were assumed to radiate as black bodies in this region of the spectrum may have an emissivity considerably less than 1; particles in the atmosphere such as aerosols or thin cirrus clouds, not visible in TV photos, may produce additional absorption not accounted for in the Wark–Yamamoto theory; the theory or assumptions used in the theory may be inadequate; or measurements given by the instrument may be in error. All four reasons are now being investigated.

An increasing deterioration in the response of all channels of the radiometer with flight time was found in Tiros II and to a greater extent in Tiros III.[15] However, corrections to account for this change in calibration are already incorporated in the temperature difference stated above. Results of further theoretical investigations and elaboration of the above reasons will certainly provide more precise methods of cloud height determinations from radiation measurements. In addition, these investigations, combined with a more complete scrutiny of all the data available, will shed light on such unknowns as the emissivity of various types of clouds and the screening effect of aerosol or similar particles on the outgoing radiation. Nevertheless, empirical correction factors can be obtained even now from existing data and successful, if approximate, determination of cloud-top heights may be and have been made from Tiros measurements.[23]

Similar considerations govern our ability to derive pertinent information on the structure of the atmosphere from radiation data. An example is given in figure 4. By comparison of the blackbody temperatures measured in the 8–12 micron window channel over hurricane Anna (200–210°K) with simultaneous measurements in the 6.7 microns water vapor absorption channel indicates that the temperatures measured in the latter are approximately 10°K higher. Since we have already determined from the window channel that the cloud tops reached to approximately 15 km (near the tropopause), this temperature difference can only be explained by the presence of considerable amounts of water vapor in the stratosphere above the cloud top.

Bandeen, et al.[9] have found by trial and error a water-vapor distribution which will produce this measured temperature difference. The required water-vapor mixing ratio vs altitude function is shown in figure 5. This function is by no means unique but was chosen because it agrees both with present ideas of water-vapor content in the stratosphere[25] and with the radiation measurements. Again, the above case is typical for

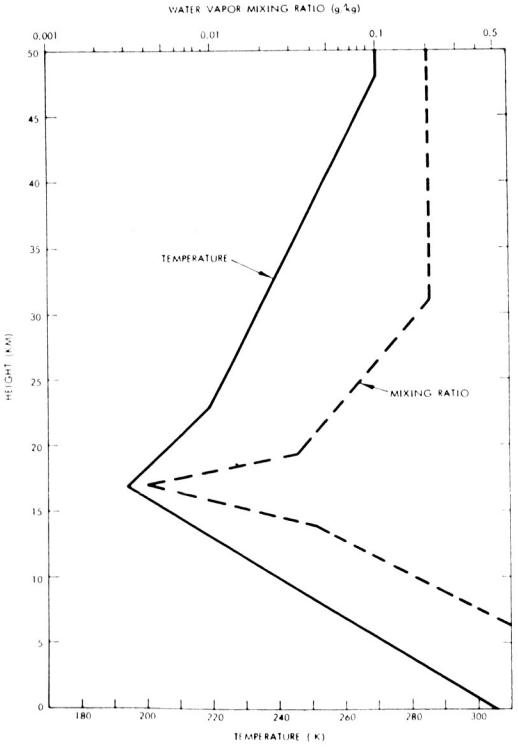


FIGURE 5.—Temperature and water vapor mixing ratio vs. height. Temperatures taken from data on tropical atmosphere. Water-vapor profile fits radiation data over Hurricane Anna.

several situations observed so far over extensive cloud systems in the tropics and is indicative of some of the more subtle information contained in these results.

Attempts to derive estimates of the global heat budget from radiation data have been made and investigations by Winston and Rao,[8] Bandeen and Nordberg,[5] Prabhakara and Rasool,[6] and House[7] have shown that on a large scale the satellite results confirm theoretical estimates by London[24] and others on the global distribution of total emitted long-wave radiation. The satellite investigations are based on Explorer VII measurements[26] and on the 8–30 micron channel from Tiros II and III.[5,6,8] An example is shown in the graph on page 1432. Tiros II data shown in this figure have been corrected for the deterioration of the sensor response with time and for total outgoing flux by the method of Wark and Yamamoto.[22]

The Tiros results show clearly, although not to the extent predicted by theory, a minimum of outgoing radiation near the equator. This minimum is due to the extensive cloud cover in the equatorial zone, while the maxima near the two tropics reflect the clear skies and warm temperatures of these regions. The more rapid decrease of outgoing radiation with latitude in the winter hemisphere, due to colder temperatures and intenser cloudiness, as well as the higher maximum near the summer tropic, can be clearly found on the satellite data. Investigation of outgoing radiation on a shorter time scale based on Tiros II

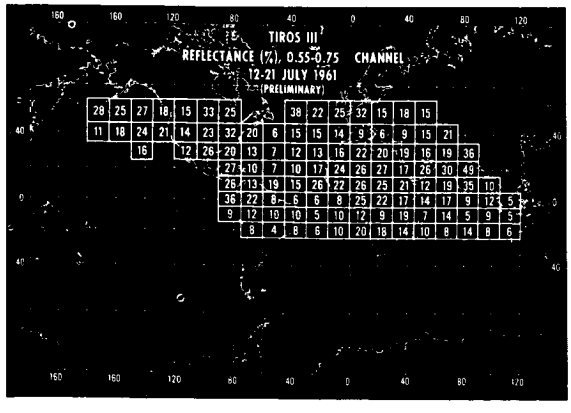


FIGURE 6.—Map of average reflected radiation in 0.55–0.75 micron channel. Numbers are in percent of incoming radiation, averaged over grid elements and time period shown.

measurements, and comparisons with dynamic considerations of the atmosphere, have shown that in some cases minima in outgoing flux go hand in hand with an increase in the zonal kinetic energy which is derived from, and is a measure of, the energy carried by winds along a latitude circle.[8] This first indication of a connection between the gross outgoing radiation measured by satellites on a planetary or continental scale and energy parameters relating to the dynamics of the atmosphere will be further pursued, and is of major importance to an understanding of the forces which create circulation in the atmosphere.

A complete planetary heat budget, however, must include a measure of incoming solar radiation in addition to outgoing long-wave radiation. Measurements from one of the Tiros III solar radiation channels (0.55–0.75 microns) of regional average of reflected energy over a period of nine days are shown in figure 6. The numbers express the total energy reflected in percent of incoming solar radiation within the spectral range of the instrument. This incoming radiation was taken to be (108.6 w/sq meter) sin$e$, where e is the elevation angle of the sun at the location and time of measurement. The value of 108.6 is derived from a total solar constant of 1395 w/sq meter and the spectral response for this channel.[16]

The illustration shows that radiometric measurements in the visible spectrum again reflect very well the distribution of cloud cover—namely, high average reflectivities near the equator and low ones in the subtropics, except for India, where during this time of the monsoon the highest average reflectances (49%) were observed. The differences in average reflectances between land and water at moderate latitudes only may also be seen in this illustration. Reflectances over the subtropical North and South Atlantic, the Indian Ocean, and the Mediterranean are appreciably lower than over land masses at comparable latitudes. Apparently, in these regions and differences in the reflectances of land and water surfaces are noticeable because of the predominantly clear weather. At higher latitudes, like the North Pacific, no difference in reflectance can be detected, apparently because cloud cover is more frequent over water than over land.

In addition to supplementing the cloud cover patterns observed during daytime in the infrared

window, which is not well suited to detect low-altitude clouds, the solar radiation channels on Tiros can be used to derive *relative* measurements of the absorption of solar energy and aid in interpretation of long-wave outgoing flux measurements.

The illustration also shows that measured reflectances cannot be interpreted as measurements of planetary albedo, and are therefore not useful to include in quantitative heat budget considerations at this time. Since the solar constant is known quite accurately and the outgoing flux is given in the graph on page 1432, we know that, for the sake of equilibrium, the sum of the average planetary albedo and the outgoing radiation must equal the incoming solar radiation. This results in a value of 32 to 36% for the average albedo. The average found from the illustration is 17 to 18%.

This disagreement by a factor of 2 is difficult to explain, particularly since the deterioration in the response of the sensors[16] has already been taken into account. The fact that Tiros measurements for this period cover only a limited region of a zone between 55°N and 25°S may influence the results somewhat, but certainly not to such a large extent.

A more likely explanation may lie in the fact that the values in the illustration are based on isotropy and uniformity with regard to both the angle and wavelength of the backscattered sunlight. There are indications from theoretical investigations[26] that the reflectance toward the direction of the incoming sunlight may be appreciably larger than in other directions. Since in most of the observations used in the illustration, the angle formed by the sun, the target, and the satellite was near 45 deg, it is possible that this might account for the low observed reflectances. A variation in the reflectance with wavelength may also contribute to make the satellite observations appear too low.

These factors are now being investigated both theoretically and by supplementary laboratory and free atmosphere balloon measurements.

CONCLUSIONS

Thus, maps of cloud patterns revealing meteorological features such as frontal systems, tropical storms, etc., have been produced with adequate resolution and in large quantities from the Tiros radiometric experiments to demonstrate their value to both day and night synoptic analysis. With appropriate, empirically derived, correction factors, one can also obtain approximate measurements of cloud heights, especially over large, uniform, and intense cloud systems. Successful determinations of the global distribution of total emitted long-wave radiation over limited periods of time have been made and agree well with theoretical predictions. Some preliminary correlations were found between outgoing fluxes and dynamic parameters in the atmosphere.

Reflected solar radiation, measured to permit an evaluation of the entire planetary heat budget, follows very closely, as one might expect, the patterns of cloudiness, as well as the distribution of land and water—at least over those regions of the globe covered by Tiros. Quantitative measurements of the earth's albedo, however, have not been possible as yet.

The difficulties which now prevent us from interpreting solar-energy reflectance measurements in terms of albedo and the disagreement found between calculated and measured emitted radiation in the window channel will certainly force us to subject measurements to further interpretations. From these we may find further enlightenment regarding the large scale scattering, emission, and transmission properties of clouds, the earth's surface, and the atmosphere—parameters which may still lie dormant in a vast amount of data. We believe that, with the successful interpretations of the radiometer measurements mentioned earlier, we have only exploited a small, more obvious portion of the data.

It has also been shown that under certain circumstances the presence of substantial amounts of water vapor in the stratosphere can be implied from simultaneous radiation measurements in two spectral channels. Again, resolution of some of the outstanding theoretical questions will produce more information on water-vapor distribution both in the troposphere and stratosphere.

Tiros radiation measurements have also pointed up the need for an improvement in the stability of optical sensors under prolonged flight environment. Deterioration in sensor response during

orbital flight has not been fully explained. This further emphasizes the need for reliable inflight calibration on future instruments. And, finally, support of satellite experiments with coordinated ground-based or balloon measurements have become even more important.

REFERENCES

1. NORDBERG, W., BANDEEN, W. R., CONRATH, B. J., KUNDE, V., and PERSANO, I.: "Preliminary Results of Radiation Measurements from the Tiros III Meteorological Satellite," *Journal of Atmospheric Sciences*, Vol. 19, No. 1, Jan. 1962, pp. 20–30, 1961.
2. BANDEEN, W. R., CONRATH, B. J., NORDBERG, W., and THOMPSON, H. P., "A Radiation View of Hurricane Anna from the Tiros III Meteorological Satellite," Space Research, *Proceedings of the Third International Space Science Symposium*, Washington, D. C., May 1962, North Holland Publishing Co., Amsterdam, 1962.
3. WEINSTEIN, LT. COL., and SOUMI, VERMER E., "Analysis of Satellite Infrared Radiation Measurements on a Synoptic Scale," *Monthly Weather Review*, U. S. Weather Bureau, 1961.
4. STAMPFL, R. A., "The Nimbus Spacecraft and Its Communication System," NASA Goddard Space Flight Center, TN D-1422, June 1962.
5. BANDEEN, W. R., and NORDBERG, W., "Summary of Reflected and Emitted Radiation Data Measured from Tiros II, III, and IV," presented at the Second Western National Meeting of the American Geophysical Union, Palo Alto, Calif., Dec. 29, 1962.
6. PRABHAKARA, C., and RASOOL, S. I., "Evaluation of Tiros Infrared Data," Space Research, *Proceedings of the Third International Space Science Symposium*, Washington, D. C., North Holland Publishing Co., Amsterdam, May 1962.
7. HOUSE, F. B., "Latitudinal Distribution of Outgoing Long-Wave Radiation Flux from Explorer VII Satellite," paper prepared at Univ. of Wisconsin, Dept. of Meteorology, Madison, Wis. To be published in *Journal of Atmospheric Sciences*, Spring 1963.
8. WINSTON, JAY S., and RAO, KRISHNA P., "Preliminary Study of Planetary-Scale Outgoing Long-Wave Radiation as Derived from Tiros II Measurements," *Monthly Weather Review*, Vol. 90, No. 8, Washington, D. C., Aug. 1962.
9. BANDEEN, W. R., KUNDE, V., NORDBERG, W., and THOMPSON, H. P., "Tiros III Meteorological Satellite Radiation Observations of a Tropical Hurricane," to be published in *Tellus*, Spring 1963.
10. CONRATH, B. J., "Earth Scan Analog Signal Relationships in the Tiros Radiation Experiment and their Application to the Problem of Horizon Sensing," NASA Goddard Space Flight Center, TN D-1341.
11. COUNTIS, THOMAS J. and YOUNG, J. B., "Space Simulation Effects on Optical Materials," Hughes Aircraft Co., TM-711, May 1962.
12. SHAH, C. K., and BARTMAN, F. L., "Discussion of the Theory of Cavity Sources for Infrared Radiation and the Calculation of Several Practical Cases," Univ. of Michigan Technical Report, Dept. of Aeronautical and Astronautical Engineering, to be published April 1963.
13. BANDEEN, W. R., and MANGER, W. P., "Angular Motion of the Spin Axis of the Tiros I Meteorological Satellite Due to Magnetic and Gravitational Torques," *Journal of Geophysical Research*, Vol. 65, No. 9, Sept. 1960.
14. HANEL, R., LICHT, J., NORDBERG, W., STAMPFL, R., and STROUD, W. G., "The Satellite Vanguard II: Cloud Cover Experiment," *IRE Transactions on Military Electronics*, MIL-4 (2 and 3) 245–247, 1960.
15. BANDEEN, W. R., HANEL, R. A., LICHT, J., STAMPFL, R. A., and STROUD, W. G., "Infrared and Reflected Solar Radiation Measurements from the Tiros II Meteorological Satellite," *Journal of Geophysical Research*, Vol. 66, No. 10, Oct. 1961.
16. *"Tiros II Radiation Data User's Manual,"* NASA Goddard Space Flight Center, Greenbelt, Md., Aug. 1962.
17. HANEL, R. A., "Low Resolution Radiometer," *ARS Journal*, Vol. 31, pp. 246–250, Feb. 1961.
18. DAVIS, J., HANEL, R. A., ET AL., "Telemetering IR Data from the Tiros II Meteorological Satellite," NASA Goddard Space Flight Center, TN D-1293, 1962.
19. *"Tiros II Radiation Data Catalog,"* NASA Goddard Space Flight Center, Greenbelt, Md., Aug. 15, 1961.
20. *"Tiros III Radiation Data Catalog,"* NASA Goddard Space Flight Center, Greenbelt, Md., Feb. 1963.
21. PARENT, R. J., MILLER, H. H., SUOMI, V. E., and SWIFT, W. B., "Instrumentation for a Thermal Radiation Budget Satellite," *Proceedings of the National Electronics Conference*, Hotel Sherman, Chicago, Ill., Oct. 12–14, 1959.
22. WARK, D. Q., YAMMAMOTO, G., and LIENESCH, J. H., "Methods of Estimating Infrared Flux and Surface Temperature from Meteorological Satellites," *Journal of Atmospheric Sciences*, Vol. 19, No. 5, Sept. 1962.
23. FRITZ, SIGMUND, and WINSTON, JAY S., "Synoptic Use of Radiation Measurements from Satellite Tiros II," reprinted from *Monthly Weather Review*, Vol. 90, pp. 1–9, Jan. 1962.
24. LONDON, J., "A Study of the Atmospheric Heat Balance," Final Report, Contract No. AF19 (122)–165, Dept. of Meteorology and Oceanography, New York Univ., 99 pp., July 1957.
25. GUTNICK, MURRAY, "Mean Annual Mid-Latitude Moisture Profiles to 31 Km," Air Force Research Report, AFCRL-62-681, July 1962.
26. DEIRMENDJIAN, D., "Scattering and Polarization Properties of Polydispersed Suspensions with Partial Absorption," Memorandum RM-3228-PR, Rand Corp., Santa Monica, Calif., July 1962.

PRELIMINARY MEASUREMENTS OF TEMPERATURES AND WINDS ABOVE 50 KM OVER WALLOPS ISLAND, VIRGINIA

WILLIAM NORDBERG AND WENDELL SMITH

Goddard Space Flight Center

Seventeen successful rocket grenade experiments were conducted at Wallops Island, Va. (38°N lat.), during the period July 1960 to June 1962. The purpose of these soundings was to describe further the atmosphere above 40 km within the framework of a previous analysis based on the IGY results at high and low latitudes. The soundings were distributed through all seasons; and some were conducted in conjunction with sodium release (above 80 km) and small meteorological rocket experiments (below 50 km). Preliminary results from the first ten grenade soundings are presented here.

Comparison of grenade with sodium results shows an abrupt change in the physical nature of the circulation pattern between 70 and 80 km. Below this altitude, the previously described seasonal circulation pattern exists while above it no regular seasonal variations can be detected. The most characteristic phenomenon for this region is the very narrow jet-stream-like band of wind at the 100 km level combined with extremely high wind shears.

The temperature structure at Wallops Island indicates the same mesospheric heating in wintertime as has been observed at Churchill. During periods of undisturbed easterly circulation, summer temperature profiles at Wallops Island exhibit the same features as typical low latitude profiles previously observed. When the easterly circulation becomes disturbed, the temperature profile becomes very similar to previously observed summertime profiles at Churchill. We find that during wintertime the atmosphere over Wallops Island still lies within the region of a strong cyclonic vortex.

INTRODUCTION

As a result of several series of rocket grenade experiments, the majority of which were conducted during IGY, attempts were made to describe the seasonal and geographic variability of the structure of the atmosphere between 30 and 90 km.[1,2,3] Some of these results were confirmed during IGY by other rocket soundings using falling sphere[4] and pressure gage techniques.[5,6] The description of the structure of the upper stratosphere and the mesophere given in Reference 1 was primarily based upon characteristic differences, derived from relatively few soundings, between IGY experiments at Churchill, Canada 59°N and pre-IGY experiments at White Sands, New Mexico 33°N. In addition, results from a small number of soundings at Woomera, Australia 31°S,[7] Johnston Island 18°N,[8] and Guam 12°N,[9] confirmed the picture. The salient features of this rather rudimentary picture were:

1. The large variation of the temperature profile in the 60 to 90 km region between high and low latitudes or between summer and winter at Churchill (figure 1) with large and multiple temperature maxima in the winter-mesosphere at Churchill (figure 2).

2. The existence of an extremely strong cyclonic circulation system up to 80 km over the entire winter hemisphere, which still prevails, much more weakly, in the equatorial zones. This vortex is replaced by anti-cyclonic circulation of lesser intensity for the summer hemisphere, again reaching far into the tropics (figure 3).

3. The observed breakdown of the wintertime

circulation up to 70 km at Churchill where meridional circulation in the stratosphere and mesosphere preceded the occurrence of a typical explosive warming at lower levels (figure 3, reference 2).

4. The systematic summer-winter variation of pressure, temperature, and density at high latitudes as opposed to low latitudes where individual variations in these parameters appear to be random.

The most surprising and least explicable of these features was the wintertime heating at the 60 to 90 km level observed at Churchill. Recent attempts by Kellogg,[9] Haurwitz,[10] and Maeda[11] to explain this phenomenon were based on these high latitude observations and on the existence of other events, such as strong subsistence at higher altitudes, events which are not susceptible to grenade measurements alone.

To substantiate this picture developed during IGY and to provide further experimental results for the analysis of the features mentioned above,

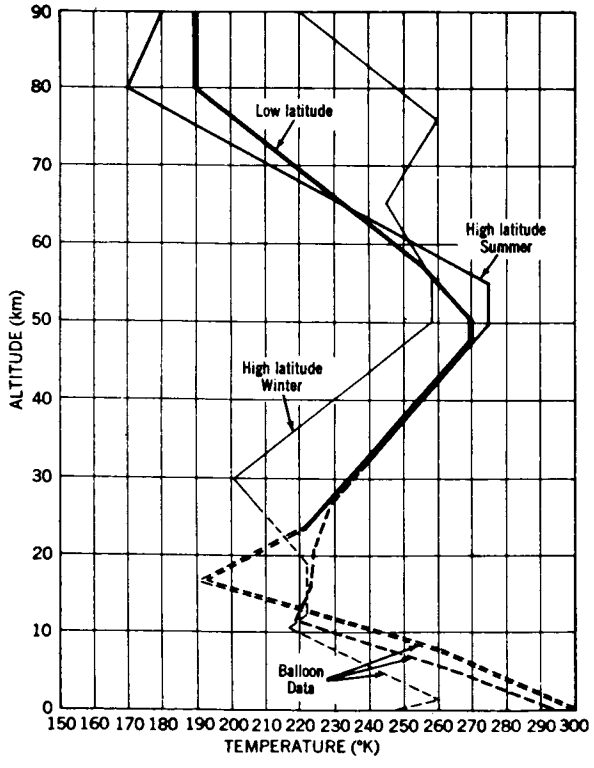


FIGURE 1.—Comparison of temperature profiles for high and low latitude atmospheres (30–90 km).

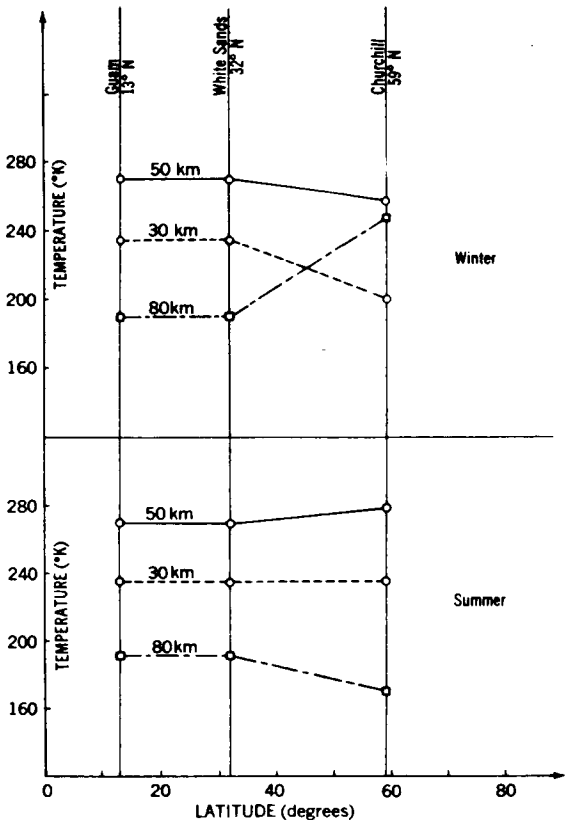


FIGURE 2.—Variation of temperatures at 30, 50, and 80 km with season and latitude.

we started in summer 1960 a series of rocket soundings at Wallops Island, Virginia 38°N. The idea was to concentrate as many soundings as possible at one single launch site and to obtain results distributed evenly over at least one full seasonal cycle. The major experiment in these soundings was again the rocket grenade experiment, but many launchings were closely coordinated with sodium releases—overlapping with and above the useful altitude range of the grenade experiment—yielding additional wind data from 60 to 200 km, and with launchings of small meteorological rockets covering the range from 25 to 50 km. Thus, by simultaneously launching sodium releases, grenade experiments, meteorological rocket dropsondes, and balloon radiosondes, a continuous wind profile was obtained from the ground up to 200 km. We intend to describe the scope of these soundings and present the preliminary results available at this time.

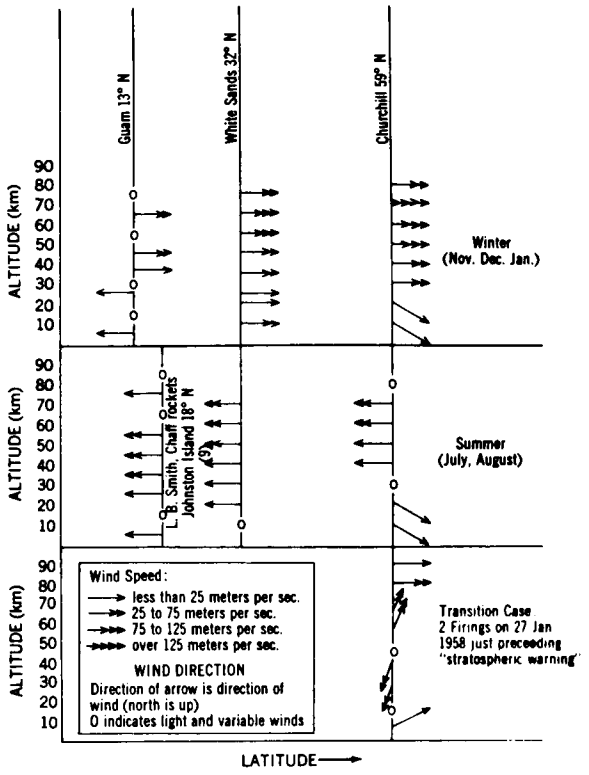


FIGURE 3.—Summary of winds up to 90 km measured at low and high latitudes.

THE ROCKET EXPERIMENTS

Balloon soundings provide measurements of the atmospheric structure parameters (pressure, density, temperature, and winds) up to about 30 km. Small, and relatively low cost, meteorological rockets of the Loki and Arcas type have the capability to carry dropsondes to 30 or 40 km above the ceiling of balloon flights. Reliable instrumentation for these rockets is limited in general, although not in principle, to the measurement of wind up to about 65 km and temperature to about 50 km. The nature of these techniques and their limitations have been reported by aufm Kampe.[12] Generally, for measurements above the mesopeak (about 50 km), more complex techniques, requiring larger and more expensive launch vehicles, must be used. The rocket grenade technique is ideally suited for the altitude range from 50 to 90 km, since it provides an accurate, simultaneous measurement of temperature and wind, and since pressure and density profiles may be derived mathematically with a high degree of precision. Detailed descriptions of the different versions of the technique have been given in references 2 and 13. Basically, it involves the determination of a temperature and wind profile from the exact measurement of the direction of sound waves arriving at the ground from a series of grenade explosions, the time and location of which must be known precisely.

At Wallops Island 12 explosive charges with their associated ejection timer, flash detectors and telemetry, plus a DOVAP tracking beacon, were carried by the Nike–Cajun vehicle, a medium range, solid propellant, two-stage sounding rocket.[14]

We found that the FPS-16 high precision tracking radar at the NASA launching site at Wallops Island can adequately skin-track the second stage Cajun rocket and determine the position of the explosions with sufficient accuracy. This system eliminates the need to carry a tracking transponder aboard the rocket. However, in the firings discussed, a transponder was carried in order to use a single station DOVAP (Doppler Velocity and Position) system to determine grenade positions as a backup to and for comparison with the FPS-16 radar tracking. The single station DOVAP determines the distance to the rocket as a function of time from one single ground transmitter-receiver station much in the same way as described previously.[2] In addition, an interferometer array of receiving antennas serves to determine the direction to the rocket. This tracking station was developed by J. Carl Seddon[15] especially for such experiments as the grenade experiment, requiring high precision tracking. Because of its simplicity, relative low cost, and high mobility it can be used in remote areas.

Aside from the availability of these tracking systems and other support facilities at Wallops Island, this launch site on the Eastern Shore of Virginia was also a good choice from a geographical point of view. The site represents a typical mid-latitude north of the subtropical (White Sands) and tropical (Guam) stations and considerably south of the subarctic (Churchill) station where the experiment had previously been performed. This series was the first opportunity

to obtain atmospheric structure data in this altitude range at mid-latitudes of the American continent.

Since the useful range of the grenade experiment is limited to altitudes of less than about 95 km,[2] many grenade firings were scheduled to coincide with sodium release experiments also conducted from Wallops Island, used the Nike–Cajun or similar launch vehicle. The method of determining wind vectors from luminescent sodium trails released by sounding rockets in the upper atmosphere had been successfully employed by Manring and Bedinger[16,17] for many years. The trails of vaporized sodium are released while the solar depression angle is near 8 degrees, so that the wind drift of the trail, still illuminated by sunlight, can be clearly photographed by three or more special tracking cameras located about

TABLE 1.—*Sodium Release and Grenade Experiments*

Rocket	Experiment	Launching time and date (EST)	Altitude range of successful data (km)
Nike Asp	Sodium release	Morning twilight, 17 Aug. 1959	140–210
Nike Asp	Sodium release	Evening twilight, 18 Nov. 1959	95–205
Nike Asp	Sodium release	Evening twilight, 24 May 1960	85–180
Nike Cajun	Sodium release	Morning twilight, 9 dec. 1960	90–140
Nike Cajun	Grenade	2259 8 July 1960	40–82
Nike Cajun	Grenade	1850 14 Feb. 1961	40–77
Nike Cajun	Grenade	2126 16 Feb. 1961	32–83
Nike Cajun	Grenade	0757 5 April 1961	42–77
Nike Asp	Sodium release	0436 19 April 1961	80–164
Nike Asp	Sodium release	1812 20 April 1961	80–180
Nike Asp	Sodium release	1439 21 April 1961	80–163
Nike Cajun	Grenade	1800 5 May 1961	39–83
Nike Cajun	Grenade	2354 5 May 1961	41–88
Nike Cajun	Grenade	1707 13 July 1961	40–92
Nike Cajun	Grenade	1102 14 July 1961	36–77
Nike Cajun	Grenade	0530 20 July 1961	34–87
Nike Asp	Sodium release	1639 16 Sept. 1961	80–208
Nike Cajun	Grenade	1855 16 Sept. 1961	31–63
Nike Asp	Sodium release	0503 17 Sept. 1961	80–175
Nike Cajun	Sodium release	1823 1 March 1962	70–134
Nike Cajun	Grenade	1901 1 March 1962	31–83
Nike Cajun	Sodium release	0554 2 March 1962	65–128
Nike Cajun	Grenade	0615 2 March 1962	38–87
Arcas	Robin sphere	1625 23 March 1962	28–48
Arcas	Radiosonde	1755 23 March 1962	28–55
Nike Cajun	Sodium release	1844 23 March 1962	58–140
Nike Cajun	Grenade	1854 23 March 1962	40–90
Nike Cajun	Sodium release	1848 27 March 1962	78–118
Nike Cajun	Grenade	1904 27 March 1962	40–93
Nike Cajun	Grenade	0428 17 April 1962	40–92
Nike Asp	Sodium release	0443 17 April 1962	77–200
Nike Cajun*	Grenade	2005 6 June 1962	40–93
Nike Asp*	Sodium release	2056 6 June 1962	60–161
Arcas*	Radiosonde	2141 6 June 1962	28–50
Arcas*	Robin sphere	2229 6 June 1962	28–55
Nike Cajun	Grenade	2053 7 June 1962	40–92

* Fired in conjunction with a pitot static tube at 1940 EST, 6 June 1962. Also, a inflatable sphere was carried by grenade vehicle and ejected at 109 km.

50–100 miles apart on the ground. Because of these severe time and weather restrictions (e.g., a clear sky is required) for the sodium experiment and because of operational difficulties at the launch site, as well as occasional rocket and payload malfunctions, it was not always possible to launch the grenade and sodium experiments simultaneously. Nevertheless, table 1 shows that in the period March 1, 1962 to June 6, 1962, there were six successful grenade soundings each launched within less than one hour of a successful sodium experiment.

For all grenade soundings, a radiosonde balloon was released within less than six hours of each grenade rocket launch. Since the number of temperature or wind measurements in each grenade firing is one less than the number of grenades exploded (for a 100 percent successful rocket, this number is 12) and since each point represents an average measurement in a layer between two explosions, it is desirable to distribute the explosions over as small an altitude range as possible. For this reason, as well as for desired comparisons, the firing of at least one small meteorological rocket, measuring winds by means of a dropsonde, was scheduled with each grenade firing. The dropsondes, usually a metallized parachute or a high drag, inflated sphere drifting with the wind, are tracked by radar and thus furnish wind information. Sometimes, the dropsondes carried thermistor temperature sensors. These dropsondes were intended to cover the altitude range from 25 to 55 km. To provide sufficient overlap, the grenade explosions were timed to go off between 40 and 95 km, leaving a layer of about 5 km between two adjacent explosions. Unfortunately, only four of the small meteorological rockets (Arcas) yielded useable data in conjunction with grenade and sodium soundings (table 1). In all other cases, either the Arcas rocket or the dropsonde payload had failed or the surface wind conditions were unsafe for the Arcas rocket to be fired.

RESULTS

All soundings listed in Table 1 were successful and produced useable data. At this time, however, results are available only from the first series of ten grenade soundings conducted during the period July 8, 1960 to September 16, 1961. Data from the remaining seven grenade soundings between March 1 and June 7, 1962 are in the process of reduction.

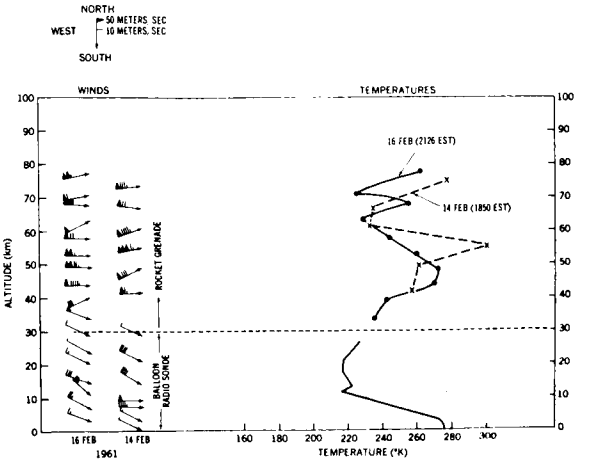


FIGURE 4.—Temperatures and winds at Wallops Island, Va., February 1961.

The results of the first ten soundings are presented in figures 4, 5 and 6. Temperature data shown in these figures are given as smooth profiles derived from up to eleven individually measured data points for each firing. Except for February 1961, the individual data points are not shown or tabulated because the intent of this paper is to demonstrate the basic characteristics of the measured temperature profiles and their relation to the previously discussed IGY results. An exception was made for the two February firings because the fine structure of the profile (due to

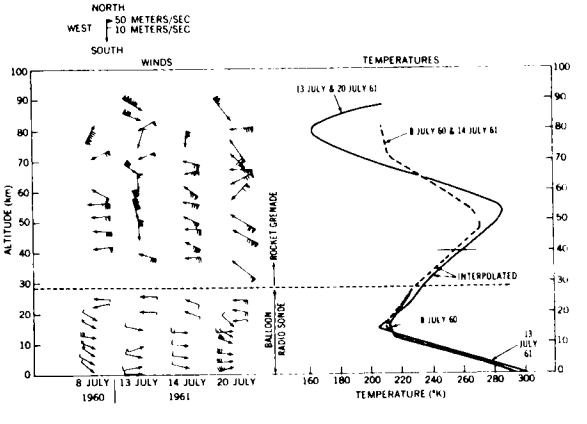


FIGURE 5.—Temperatures and winds, Wallops Island, Va., July 1960, July 1961.

the individually measured points) is so significant that it warrants presentation even in this general study. A full tabulation of all data, including a detailed error analysis, is now in progress and will be published at a later date.

The wind data in figures 4, 5 and 6 represent measured average values for the altitude interval between two grenade explosions. In general, the altitude of the explosion lies in the middle between two data points shown. Wind data from the 14 sodium releases prior to April 17, 1962 have been reduced and reported by Manring.[16,17] Two typical

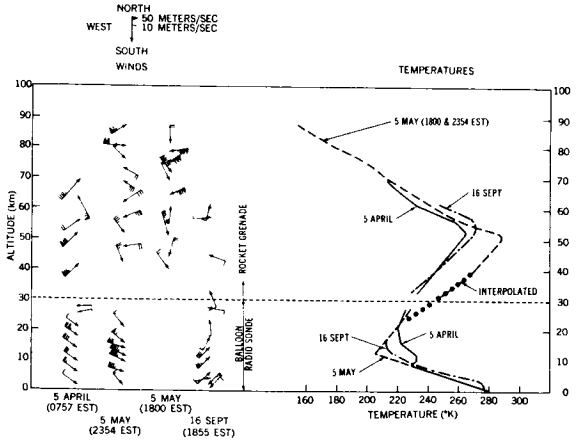


FIGURE 6.—Temperatures and winds, Wallops Island, Va., spring and fall 1961.

wind profiles are reproduced in figures 7 and 8; the two releases were conducted during the night of 1/2 March 1962, one at twilight, the other at dawn. A series of significant features is common to most of the 14 profiles studied—between about 80 to 90 km there is a sharp and rapid transition in the wind regime. Below this altitude, the commonly known pattern of generally uniform zonal flow, regularly reversing with season, exists. Above, there is a region where strong but highly variable winds are sandwiched between zones of relative calm. This region is characterized by the existence of remarkable wind shears. Thus far, every sounding conducted has evidenced these wind shears between 90 and 110 km. A typical case is shown in figure 8 where, within an interval of less than 5 km, the wind speed increases by more than 100 m/sec. This jet-stream-like band of maximum wind velocity is usually located between

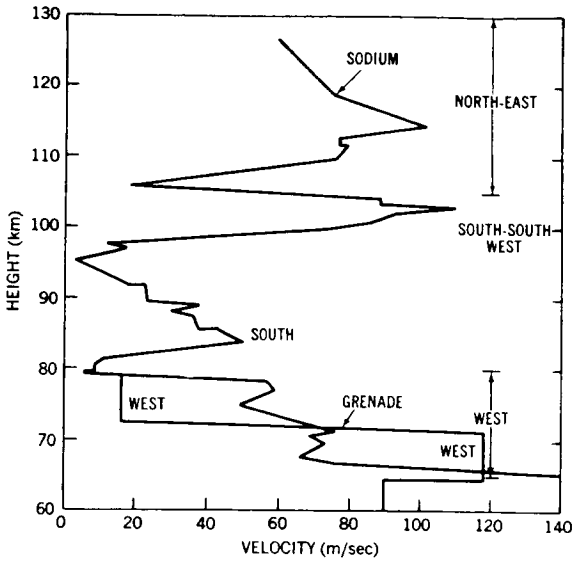


FIGURE 7.—Wind speeds and directions from the sodium release experiment, Wallops Island, Va., sunrise March 2, 1962.

95 and 105 km. Immediately below and above the wind velocities diminish almost to zero. Above 110 km, there is again a region of strong winds which may extend to altitudes as high as 200 km; however, as can be seen from table 1, very few soundings reached above 170 km. So far, we have been unable to derive any clear cut patterns for the wind directions from these 14 firings. There is

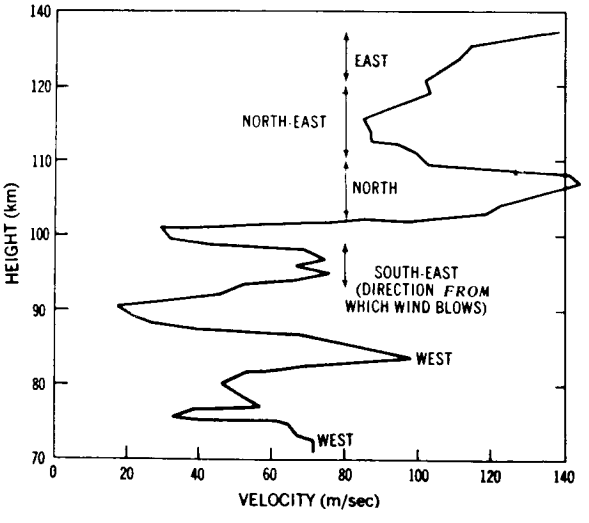


FIGURE 8.—Wind speeds and directions from the sodium release experiment, Wallops Island, Va., sunset March 1, 1962.

no definite variation either with season or between evening and morning soundings. In some but not all soundings, the southerly wind component (winds from the south) seems to prevail for the wind speed maximum between 95 and 105 km. Above 110 km, a strong component from the north is usually found.

In figure 7, the winds derived from the sodium release are compared with preliminary wind results from a nearly simultaneous grenade sounding. Considering that the grenade winds represent averages over a relatively wide altitude range (6–8 km) and that the wind speed is rapidly changing with altitude, we find the agreement between the two methods quite satisfactory.

Below 80 km, the grenade wind results from Wallops Island show a pattern which confirms the previous descriptions of circulation at these altitudes. Both February 1961 soundings indicate very strong winds from the west. There are no significant meridional wind components at any altitude. A region of maximum wind speeds seems to exist between 50 and 60 km where west winds of 125 and 165 m/sec were observed on February 16 and February 14, respectively. These values are in accord with those observed at the same altitude at Churchill on November 12, 1956.[2]

The four summer soundings shown in figure 5 exhibit a very interesting pattern. Two wind profiles for July 8, 1960 and July 14, 1961 again confirm the easterly circulation previously found over the summer hemisphere. Generally uniform flow from the east prevails up to 70 km. Wind speeds are much lower than in winter, with maxima not over 75 m/sec. This and the fact that the prevailing easterly flow ceases at about 70 km, are again in agreement with the Churchill observations for July and August 1957.[2] The two soundings on July 13 and July 20, 1961, however, depart markedly from this pattern. Both show very strong meridional components. In fact, on July 13 strong winds from almost due north prevail between 45 and 60 km. On July 20, again there is a significant meridional component, although the zonal component is always from the east. Up to 55 km, the meridional component is from the south, while between 55 and 65 km, it shifts to the north. On both days, the wind speeds are unusually high (maxima between 100 and 150 m/sec) for summer conditions. In all four cases, the winds above 70 km are quite strong and irregular in direction.

Results from April 5, 1961 indicate generally a westerly component of the zonal flow, although the wind speeds have diminished appreciably when compared to the February soundings, and a meridional component from the south of almost the same strength as the westerly component is noticeable (figure 6). An exception can be found around 60 km, where a strong south-east wind (65 m/sec) was measured. On May 5, as evidenced by both soundings on that date, the westerly flow has broken down completely. Very strong wind shears exist at all altitudes, and, although at some interspersed levels an easterly component prevails, there is no indication of the easterly summer circulation. In fact, only one out of the five soundings conducted between May 5 and July 20, 1961 showed the typical summertime flow (figure 5).

Much in contrast to the breakdown of the westerly circulation in springtime where winds were still very strong, but not uniform in direction, the results from the sounding on September 16, 1961 indicate extremely weak winds over the whole altitude range sampled (figure 6). Again, this is in agreement with the Churchill data,[2] where relatively strong meridional flow was observed during the breakdown of the winter circulation on January 27, 1958, but very weak winds existed late in August 1957 during the transition from summer to winter flow.

Several features of the temperature results are most interesting. The extremely high temperatures above 50 km in the two February soundings, with positive and negative lapse rates rapidly alternating, is very reminiscent of the Churchill winter mesosphere.[2] This heating, now observed at Wallops Island, was previously believed to exist only at high latitudes. The very high temperatures (300° and 280°K) observed on February 14 at about 55 and 75 km, respectively, seem to have penetrated to somewhat lower levels (48 and 68 km) and diminished substantially (275° and 257° K) by February 16 (figure 4).

The four temperature profiles measured during summer should be considered in conjunction with

the wind profiles. The two cases for which a uniform easterly flow prevailed (July 8, 1960 and July 14, 1961) show a typical "low latitude" temperature profile.[1] The maximum ranges around 275°K and lies just below 50 km. At 80 km, a relatively shallow temperature minimum of about 200°K to 210°K is indicated. In contrast, the two soundings on July 13 and 20, 1961, where strong meridional wind components were observed, show entirely different temperature profiles. They are characterized by a high temperature maximum (about 280°–290°K) just above 50 km and a very steep temperature minimum of about 160°K at 80 km (figure 5).

On May 5, 1961, in both soundings we again have a very similar situation. A rather disturbed wind flow is accompanied by a high mesopeak (285°K) and a low mesopause temperature (160°K at 85 km). In contrast, the sounding of April 5 indicates a mesopeak of only about 265°K. The high temperatures above 50 km observed in February have disappeared.

The temperature profile for September 16 exhibits the same structure as the April sounding. Unfortunately, both the April 5 and the September 16 soundings did not yield any data above 70 km.

CONCLUSIONS

The wind pattern observed over Wallops Island agrees, in principle, with our previous analysis based on the Churchill, White Sands, Guam, and other, low latitude soundings, and with the circulation schemes derived by Murgatroyd[18] and Batten.[19] From the coordination between sodium and grenade soundings, one may definitely conclude that these circulation patterns, most likely based upon the distribution of temperature and reflecting the meteorological behavior of this portion of the atmosphere, cease to exist above 70 or 80 km. Above this level, other factors, probably tidal forces, become increasingly prevalent in determining the circulation which does not follow the established seasonal pattern which exist at lower altitudes. Since only few of the ten grenade soundings reported here have reached above 80 km and since the sodium technique is restricted to twilight periods only, it is impossible on this basis to reach a definite conclusion in regard to this driving force behind the strong and variable wind systems above 80 km. In explaining the remarkably persisting E-region jet-stream between 95 and 105 km with its associated high shears, we will probably have to abandon the meteorologist's methods dealing with the neutral atmosphere of constant composition at lower altitudes and use those peculiar to the electrically charged ionosphere.

We conclude that, at least on February 14 and 16, 1961, the circumpolar vortex characteristic from wintertime high latitudes reached as far south as Wallops Island 38°N without any appreciable loss in strength. From these, as well as from the December 1957 results at Churchill, we conclude that the large and sporadic temperature peaks between 50 and 80 km in winter are not necessarily connected with breakdowns of the wind pattern, but are found when the established westerly circulation is very strong. The February temperature results also indicate that previous theories, explaining the high temperatures between 50 and 80 km only for high latitudes, must be revised.

The July results offer the very interesting conclusion that in summer at mid-latitudes, we may find two types of temperature distributions in rapid succession: When the circulation is uniformly from the east, the temperatures up to 80 km are in accord with those prevailing at low latitudes. During periods of severe disturbances in the easterly flow, the temperature profile changes to one typical for high latitude summers. The transition between the two types of profiles can take place within 18 hours, as evidenced by the two soundings on July 13 and 14. The conclusion that high temperatures at about 50 km and below and very cold temperatures at the mesopause are associated with a strongly meridional circulation may also be drawn from the Spring 1961 results. The September 1961 sounding seems to suggest that the change from the low latitude to the high latitude temperature profile is related to the presence of a strong meridional flow rather than to the absence of easterly circulation. The previous conclusion that this strong meridional flow is more likely to exist during the spring rather than the autumn transition period is also confirmed.

ACKNOWLEDGMENTS

We greatly appreciate the contributions of Dr. Harold Allen and Mr. Wallace Wilkie in designing and constructing the payload instrumentation for the grenade experiment, and of Mr. Alan Dean and Mr. James Pruitt in conducting the sound ranging portion of this experiment.

REFERENCES

1. NORDBERG, W., and STROUD, W. G., "Seasonal, Latitudinal and Diurnal Variations in the Upper Atmosphere," NASA Technical Note D-703, April 1961.
2. STROUD, W. G. NORDBERG, W., BANDEEN, W. R., et al., "Rocket-Grenade Measurements of Temperatures and Winds in the Mesosphere over Churchill, Canada," *J. Geophys. Res.* **65**(8): 2307–2323, August 1960.
3. NORDBERG, W., and STROUD, W. G., "Results of IGY Rocket-Grenade Experiments to Measure Temperatures and Winds above the Island of Guam," *J. Geophys. Res.* **66**(2): 455–464, February 1961.
4. JONES, L. M., PETERSON, et al., "Upper-Air Density and Temperature: Some Variations and an Abrupt Warming in the Mesosphere," *J. Geophys. Res.* **64**(12): 2331–2340, December 1959.
5. AINSWORTH, J. E., FOX, D. F., and LAGOW, H. E., "Upper-Atmosphere Structure Measurement Made with the Pitot-Static Tube," *J. Geophys. Res.* **66**(10): 3191–3212, October 1961.
6. SPENCER, N. W., "Density and Temperature in the Mesosphere above Churchill, Canada," (Submitted to *J. Geophys. Res.*).
7. GROVES, G. V., and others, "Wind and Temperature Results Obtained in Skylark Experiments," in: *Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January 1960,* Amsterdam: North-Holland Publ. Co., pp. 144–153., 1960.
8. SMITH, L. B., "The Measurement of Winds Between 100,000 and 300,000 ft by use of Chaff Rockets," *J. Meteorol.* **17**(3): 296–310, June 1960.
9. KELLOGG, W. W., "Chemical Heating above the Polar Mesopause in Winter," *J. Meteorol.* **18**(3): 373–381, June 1961.
10. HAURWITZ, B., "Frictional Effects and the Meridional Circulation in the Mesosphere," *J. Geophys. Res.* **66**(8): 2381–2391, August 1961.
11. MAEDA, K., "On the Heating of the Polar Upper Atmosphere," NASA Technical Report R-141, 1962.
12. AUFM KAMPE, H., "Review of U. S. Meteorological Rocket Network Activities and Results," in: *Proc. 1st Internat. Sympos. on Rocket and Satellite Meteorology, Washington, April 1962,* Amsterdam: North-Holland Publ. Co. (To be published).
13. STROUD, W. G., TERHUNE, E. A., et al., "Instrumentation of the Rocket-Grenade Experiment for Measuring Atmospheric Temperatures and Winds," *Rev. Sci. Instr.* **26**(5): 427–432, May 1955.
14. NORDBERG, W., and SMITH, W. S., "The Rocket Grenade Experiment," (Proposed NASA Technical Note).
15. SEDDON, J. C., "Preliminary Report on the Single Station Doppler-Interferometer Rocket Tracking Technique," NASA Technical Note D-1344, 1962.
16. MANRING, E., BEDINGER, J., KNAFLICH, H., and LYNCH, R., "Upper Atmosphere Wind Profiles Determined from Three Rockets Experiments," Geophysics Corporation of American Technical Report 61-1-N, February 1961.
17. MANRING, E. R., "Study of Winds, Diffusion, and Expansion of Gases in the Upper Atmosphere," Quarterly Progress Report, 1962, Submitted by Geophysics Corporation of America Under NASA/GSFC Contract No. NASW-396.
18. MURGATROYD, R. J., "Winds and Temperatures Between 20 km and 100 km—A Review," *Quart. J. Royal Meteorol. Soc.* **83**(358): 417–458, October 1957.
19. BATTEN, E. S., "Wind System in the Mesosphere and the Lower Mesosphere," *J. Meteorology* **18**(3): 283–291, June 1961.

ROCKET MEASURMENT OF THE STRUCTURE OF THE UPPER STRATOSPHERE AND MESOSPHERE

W. NORDBERG AND W. SMITH

Goddard Space Flight Center

INTRODUCTION

The region between 30 and 200 km is often referred to as the forgotten region of the atmosphere or the "ignosphere" and our ignorance of the detailed physical processes in this region is attributed to the fact that it lies beyond the reach of routine balloon experiments and is not well suited for exploration by orbiting spacecraft. One might consider, however, that even less information exists about the "satellite" regions above 200 km and that the few measurements made so far in that region provide convenient "anchor points" for a crude understanding of the very basic atmospheric properties. At the lower altitudes, especially in the 30 to 100 km region we have already passed this first step of understanding the basic properties and have entered the next and perhaps much more difficult phase of describing and explaining the "synoptic" of behaviour of these properties. Probing this portion of the atmosphere with various types of rocket sondes over the past decade has produced a sufficient amount of data to permit at least an attempt to describe the variation of the temperature structure and the circulation in this region on a global scale. Such descriptions may form the basis for further analyses of the physical processes governing the observed patterns of atmospheric structure and provide answers to the question of interaction between solar energy and the state of the atmosphere. It should be the primary object of theory and experimentation to explore this question and to obtain an understanding of the dynamics of this region which lies between the directly solar influenced thermosphere and the troposphere where meterological processes often seem to occur randomly. In other words our aim should be to establish the foundations of a "meteorology" of the upper stratosphere and mesophere.

REVIEW OF ROCKET RESULTS THROUGH IGY

As a result of several series of rocket grenade experiments, the majority of which were conducted during IGY, attempts were made to describe the seasonal and geographic variability of the structure of the atmosphere between 30 and 90 km.[1–3] Some of these results were confirmed during IGY by other soundings using falling sphere[4] and pressure gage techniques.[5,6] The description of the structure of the upper stratosphere and mesophere given in reference 1 was primarily based upon characteristic differences, derived from relatively few soundings, between IGY experiments at Churchill, Canada, 59°N and pre-IGY experiments at White Sands, New Mexico, 33°N. In addition, results from a small number of soundings at Woomera, Australia, 31°S,[7] Johnston Island, 18°N,[9] and Guam, 12°N,[3] confirmed the picture. The salient features of this rather rudimentary picture are:

1. The large variation of the temperature profile in the 60 to 90 km region between high and low latitudes or between summer and winter at Churchill with large and multiple temperature maxima in the winter mesosphere at Churchill.

2. The existence of an extremely strong cyclonic circulation system up to 80 km over the entire winter hemisphere, which still prevails, much

more weakly, in the equatorial zone. This vortex is replaced by anti-cyclonic circulation of [illegible] intensity for the summer hemisphere, again reaching far into the tropics.

3. The observed breakdown of the wintertime circulation up to 70 km at Churchill where meridional circulation in the stratosphere and mesosphere preceded the occurrence of a typical explosive warming at lower levels.

4. The systematic seasonal variation of pressure, temperature, and density at high latitudes where variations by a factor of two in density were observed between summer and winter at 60 km.

The most surprising and least explicable of these features was the wintertime heating at the 60 to 90 km level observed at Churchill. Recent attempts by Kellogg,[9] Haurwitz[10] and Maeda[11] to explain this phenomenon were based on these high altitude observations and on the existence of other events, such as strong subsidence at higher altitudes, which events are not susceptible to measurements by present rocket probes.

PRESENT ROCKET PROGRAMS

To substantiate this picture developed during IGY and to provide further experimental results for the analysis of the features mentioned above, we started, in summer 1960, a series of rocket soundings at Wallops Island, Virginia, 38°N. This launch site on the Eastern Shore of Virginia provides both excellent support and tracking facilities and is also a good choice from a geographical point of view. The site represents a typical mid-latitude north of the subtropical (White Sands) and tropical (Guam) stations and considerably south of the subarctic (Churchill) station were the experiment had previously been performed. This series was the first opportunity to obtain atmospheric structure data in this altitude range at mid-latitudes of the American continent. The idea was to concentrate as many soundings as possible at one single launch site and to obtain temperature and wind profiles distributed evenly over at least one full seasonal cycle. The major experiment in these soundings was again the rocket grenade experiment, but many launchings were closely coordinated with sodium releases—overlapping with and above the useful altitude range of the grenade experiment—yielding additional wind data from 60 to 200 km, and with launchings of small meteorological rockets covering the range from 25 to 50 km. Thus, by simultaneously launching sodium releases, grenade experiments, meteorological rocket dropsondes, and balloon radiosondes, a continuous wind profile was obtained from the ground up to 200 km.

In addition to this series of soundings the results of which will be described below, there exists a U.S. Meteorological Rocket Network using small (5 to 10 lb payload) and inexpensive ($ 2,000) meteorological rockets. This network performs regular soundings (several times per month) from six or more stations over the North American continent. The soundings provide good wind profiles and occasional temperature measurements throughout the stratosphere.[12]

There are definite plans to conduct the grenade and the sodium experiments, which require larger (50–80 lb) payloads and more expensive ($ 10,000–25,000) rockets simultaneously at Churchill, Canada and at Wallops Island, Virginia. It is hoped that these and similar experiments can be extended to other geographic regions and that together with soundings already in progress in Australia and Japan a truly global exploration of the stratosphere and mesosphere can be accomplished. Such an accomplishment during the forthcoming IQSY would provide an appropriate climax to an effort so successfully initiated during IGY.

RESULTS OF RECENT ROCKET SOUNDINGS AT WALLOPS ISLAND, VIRGINIA

All successful soundings are listed in table 1. The resulting preliminary data are presented in figures 1 through 11. Each wind and temperature point resulting from the grenade experiment and presented in figures 1 through 11 constitutes a measured average value for the altitude interval between two grenade explosions. In general, the altitude of an explosion lies in the middle between two data points shown. The winds measured by the Grenade Experiment are supplemented, in many cases, by wind measurements with the sodium release technique, and by meteorological rocket dropsondes to study the transition of the circulation from the mesosphere to the ionosphere,

TABLE 1.—*Sodium Release and Grenade Experiments*

Rocket	Experiment	Launching time and date (EST)	Altitude range of successful data (km)
Nike Asp	Sodium release	Morning twilight, 17 Aug. 1959	140–210
Nike Asp	Sodium release	Evening twilight, 18 Nov. 1959	95–205
Nike Asp	Sodium release	Evening twilight, 24 May 1960	85–180
Nike Cajun	Sodium release	Morning twilight, 9 Dec. 1960	90–140
Nike Cajun	Grenade	2259 8 July 1960	40–82
Nike Cajun	Grenade	1850 14 Feb. 1961	40–77
Nike Cajun	Grenade	2126 16 Feb. 1961	32–83
Nike Cajun	Grenade	0757 5 April 1961	42–77
Nike Asp	Sodium release	0436 19 April 1961	80–164
Nike Asp	Sodium release	1812 20 April 1961	80–180
Nike Asp	Sodium release	1439 21 April 1961	80–163
Arcas	Radiosonde	0953 5 May 1961	21.5–44.5
Nike Cajun	Grenade	1800 5 May 1961	39–83
Nike Cajun	Grenade	2354 5 May 1961	41–88
Nike Cajun	Grenade	1707 13 July 1961	40–92
Nike Cajun	Grenade	1102 14 July 1961	36–77
Nike Cajun	Grenade	0530 20 July 1961	34–87
Nike Asp	Sodium release	1639 16 Sept. 1961	80–208
Nike Cajun	Grenade	1855 16 Sept. 1961	31–63
Nike Asp	Sodium release	0503 17 Sept. 1961	80–175
Nike Cajun	Sodium release	1823 1 March 1962	70–134
Nike Cajun	Grenade	1901 1 March 1962	31–83
Nike Cajun	Sodium release	0554 2 March 1962	65–128
Nike Cajun	Grenade	0615 2 March 1962	38–87
Arcas	Robin sphere	1625 23 March 1962	28–48
Arcas	Radiosonde	1755 23 March 1962	28–55
Nike Cajun	Sodium release	1844 23 March 1962	58–140
Nike Cajun	Grenade	1854 23 March 1962	40–90
Arcas	Radiosonde	1800 27 March 1962	28–55
Nike Cajun	Sodium release	1848 27 March 1962	78–118
Nike Cajun	Grenade	1904 27 March 1962	40–93
Nike Cajun	Grenade	0428 17 April 1962	40–92
Nike Asp	Sodium release	0443 17 April 1962	77–200
Nike Cajun†	Grenade	2005 6 June 1962	40–93
Nike Asp†	Sodium release	2056 6 June 1962	60–161
Arcas†	Radiosonde	2141 6 June 1962	28–50
Nike Cajun	Grenade	2053 7 June 1962	40–92

† Fired in conjunction with a pitot static tube at 1940 EST, 6 June 1962. Also, an inflatable sphere was carried by grenade vehicle and ejected at 109 km.

and to provide comparisons in the regions of overlap respectively.

In three cases, grenade temperature measurements are compared to nearly simultaneous meteorological dropsoundings. Temperature comparisons, are shown in figures 10 and 11 and wind comparisons between the small meteorological rocket dropsondes and the grenade techniques are shown in figure 6. The wind data from the meteorological soundings were obtained with both Arcas Robin (small inflated falling sphere), and Arcas parachute dropsondes, (reference 13). There is good agreement in the winds in all cases. The temperatures resulting from Arcas thermistor dropsondes[12] seem to compare with the grenade temperatures only at altitudes below 4 km and

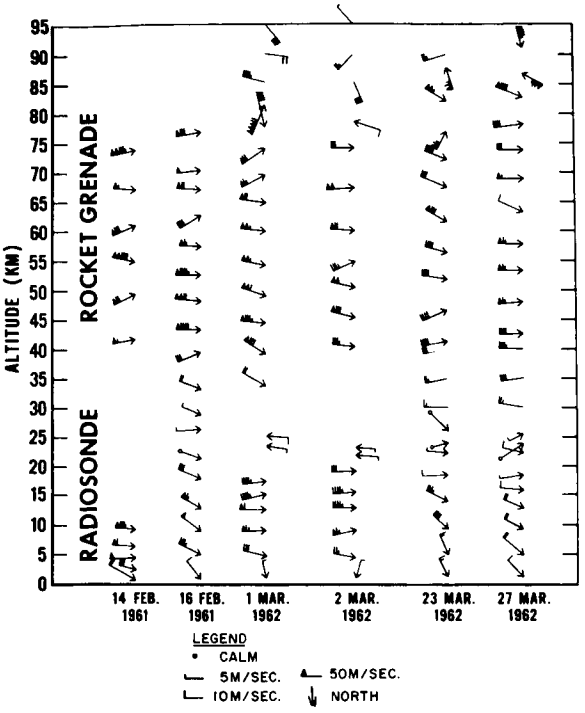


FIGURE 1.—1961–62 winter winds up to 95 km at Wallops Island, Va. (38°N). Wind vectors shown with arrows were obtained from grenade experiment. Wind vectors without arrows stem from Arcas-sondes (below 60 km) or from sodium release (above 70 km).

there, only in the case of 5 May 1961. Above 40 km the dropsonde temperatures are about 5 to 10°K higher than the grenade temperatures. This discrepancy could possibly still be within the error limit of the two experiments, or could be due to the relatively long time interval between the Arcas and grenade soundings. However, on 23 and 27 March 1962, the disagreement is considerably worse (figure 10). At 45 km temperatures from the small meteorological rockets (Arcsondes) are 15 to 20°K higher than grenade temperatures, and at 60 km where this difference has increased to 45°K, the Arcsonde temperatures are apparently meaningless; we have learned that the two Arcsondes in March 1962 carried thermistor temperature sensors, which were particularly sensitive to heating by RF radiation from the telemetering transmitter of the sonde. The production of this type of sensor has since then been discontinued. Unfortunately only three temperatures comparisons were possible because in all all other attempts the Arcas rocket or the temperature sensors failed or high surface wind made the launching of the Arcas unsafe.

Figures 1 through 3 demonstrate well the seasonal nature of the circulation above Wallops Island, Virginia (38°N) between the altitudes of 30 and 80 km. The flow is primarily zonal with prevailing westerly winds during the winter and prevailing easterly winds during the summer season. These results support substantially the pattern derived previously;[1] namely, the predominance throughout the winter mesosphere and stratosphere of an intense circumpolar vortex whose strength decreases toward the equator and a similar, somewhat weaker anticyclonic circulation in the same region in summer.

Figures 1 through 3 clearly exhibit the transition from the highly uniform and strong westerly flow in winter to the easterly flow in summertime. The winds in late March (figure 1) and early April (figure 3) show a somewhat lower velocity

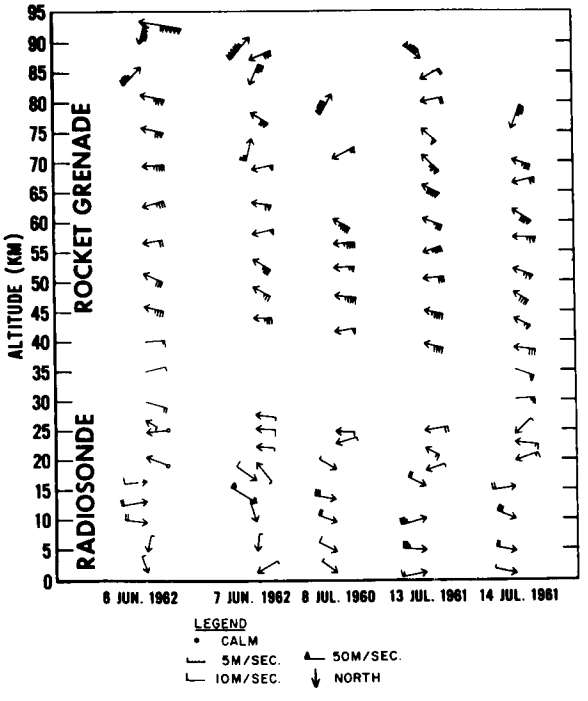


FIGURE 2.—1960, 1961, 1962 summer winds up to 95 km at Wallops Island, Va. (38°N). Wind vectors shown with arrows were obtained from grenade experiment. Wind vectors without arrows stem from Arcas-sondes (below 60 km) or from sodium release (above 70 km).

and increasing variability as the year progresses. The flow on 5 May shows practically no zonal component in the stratosphere. In this region the winds are predominantly from the north while in the mesosphere the circulation already seems to have reversed to an easterly flow. It is interesting to observe that on this day the easterly winds in the 75 to 80 km region are stronger than in any sounding during the summer. Throughout the 40 to 70 km region the 5 May soundings exhibit more variability in the wind structure than any observations on other days of the year. It is also interesting to note that the wind vectors in the 55 to 60 km and 70 to 75 km regions have undergone tremendous changes in the less than 6 hours between the two soundings. As will be discussed

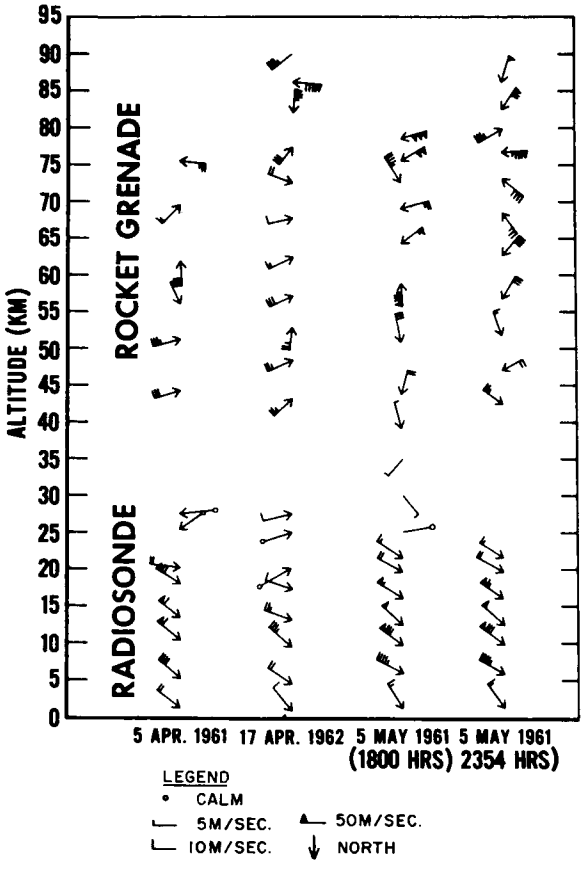


FIGURE 3.—Winter to summer—transition period winds up to 95 km at Wallops Island, Va. (38°N), April 1962, May 1961—wind vectors shown with arrows were obtained from grenade experiment. Wind vectors without arrows stem from Arcas-sondes (below 60 km) or from sodium release (above 70 km).

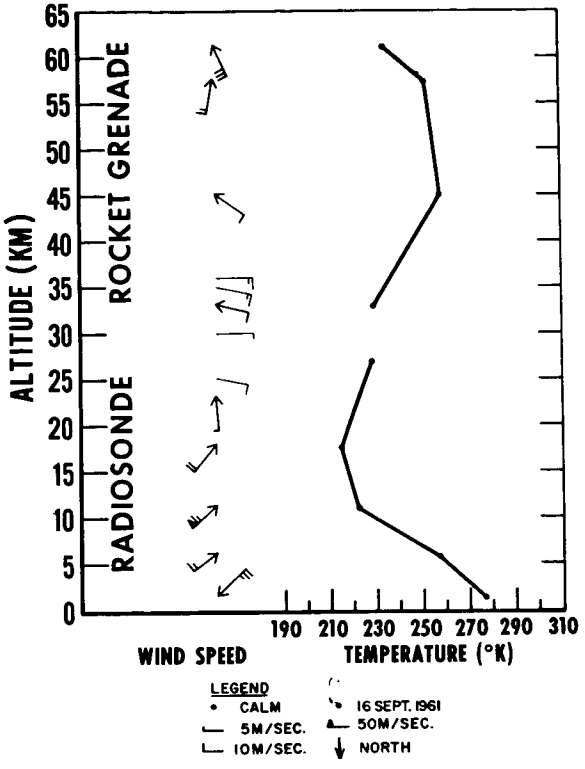


FIGURE 4.—Temperatures and winds up to 65 km at Wallops Island, Va. (38°N), September 1961.

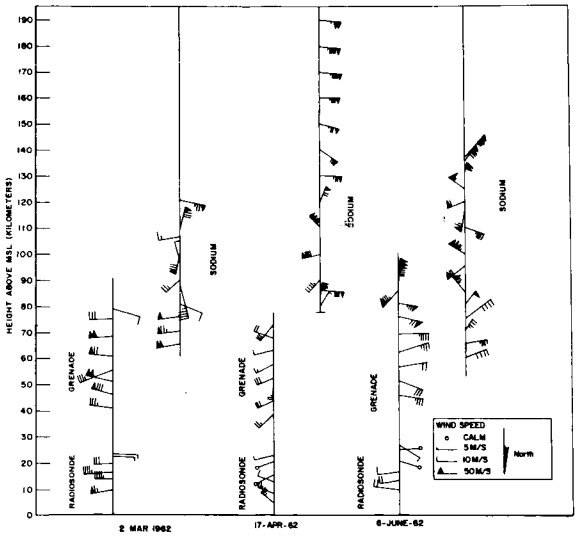


FIGURE 5.—Comparison of winds at Wallops Island (38°N) measured by sodium and grenade technique March, April, and June 1962. Sodium winds shown represent *instantaneous* values measured or interpolated at altitude shown, grenade winds represent *average* winds for the altitude interval between two grenade explosions.

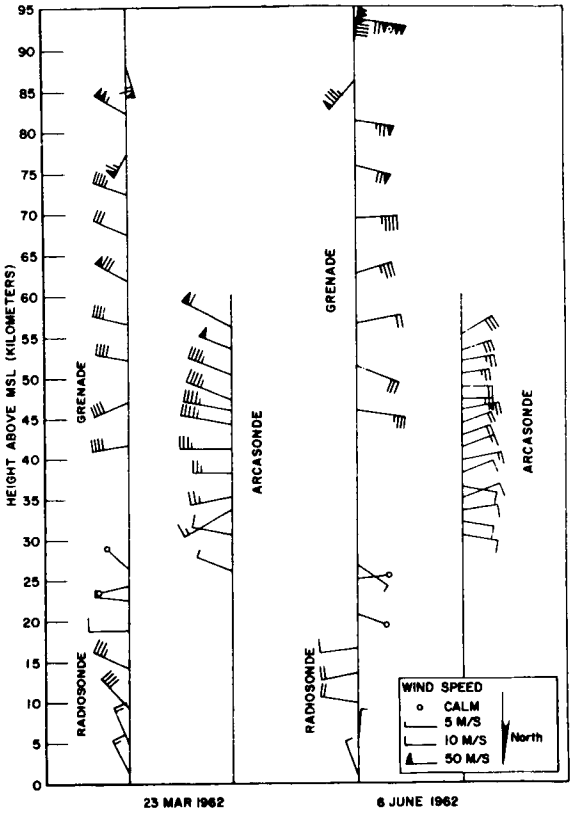


FIGURE 6.—Comparison of winds at Wallops Island (38°N) measured by Arcas and grenade techniques. March and June 1962.

below the stratopause temperature has also reached a maximum on this day.

Weak winds were again observed in mid-September (figure 4). This is in good agreement with 1957 Fort Churchill soundings, which indicated the beginning of the transition to winter circulation during late August.[2]

Most of the wind measurements above 80 km were obtained by use of the sodium release experiment and preliminary results have been reported by Manring.[14] The grenade technique, which yields average winds in layers of a few kilometers thickness up to 95 km, already indicates the variability of the wind above 80 km. The sodium measurements, yielding a continuous wind profile, give a more adequate description of the remarkable shears, characteristic of the 80 to 120 km region (figure 5). Where the winds are uniform with season (below 80 km), a better basis exists for comparing the average grenade measurements with the instantaneous sodium wind track. Such comparison between the two methods can be seen in figure 5. In most cases the agreement is fair.

A sharp and remarkable boundary seems to separate the circulation in the mesosphere from the circulation in the lower ionosphere. This boundary lies near 80 km and seems to suggest that the physical causes which drive the atmosphere are quite different in the two regions.

As described above, the winds below 80 km conform to the pattern of uniform zonal flow regularly, reversing with season, interrupted only by occasional breakdowns during the spring transition. Above this altitude the flow is no longer uniform and exhibits no regular seasonal pattern. Some features are common, nevertheless, to most of the wind profiles taken above 80 km;

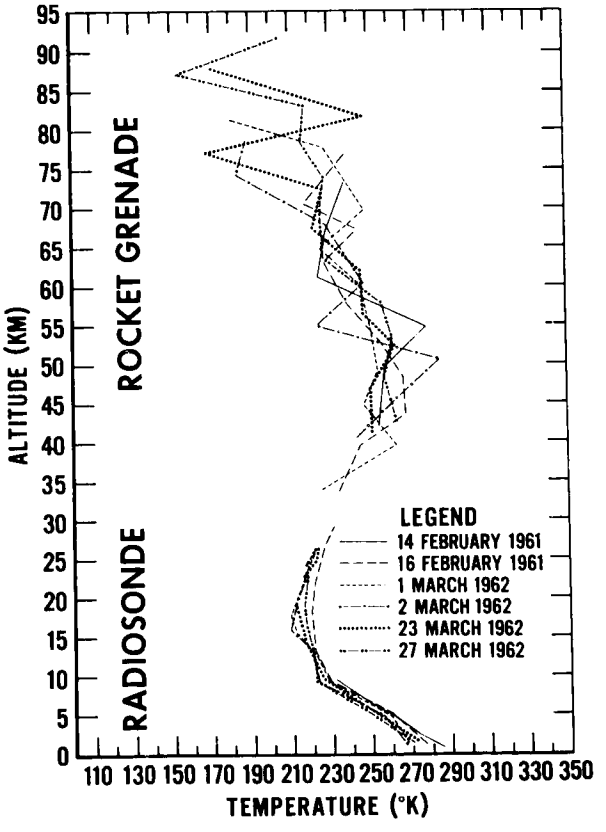


FIGURE 7.—1961, 1962 winter temperatures up to 95 km at Wallops Island, Va. (38°N).

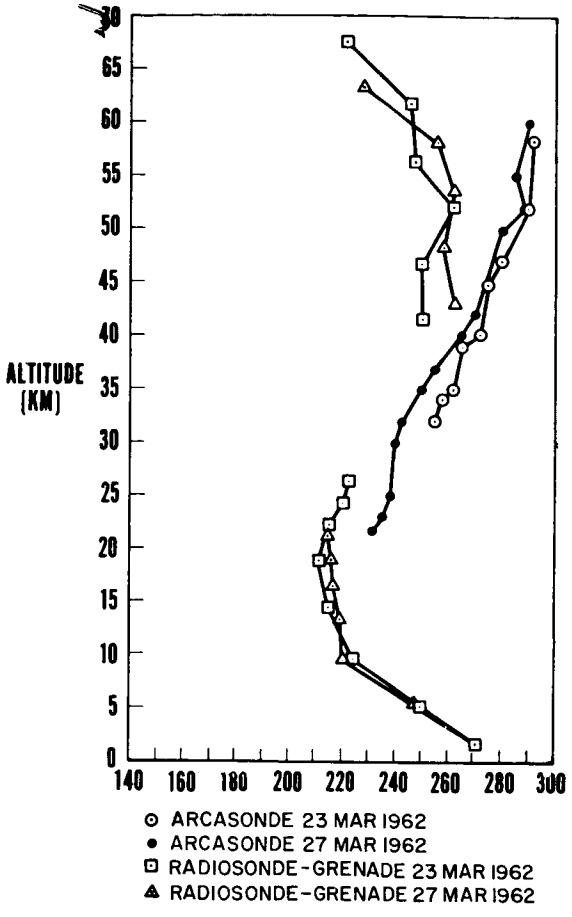


FIGURE 10.—Comparison of temperatures at Wallops Island (38°N) measured by Arcas and grenade technique, March 1962.

The temperature profiles during the transition from a winter regime to a summer regime are also of interest (figure 9). By April the source of heating at the 80 km region has disappeared resulting in a steeper mesodecline, characteristic of a summertime condition.

The midnight sounding on 5 May 1961, however, exhibits the fluctuations of a typical wintertime profile, much in contrast to the soundings six hours before as well as to the two April 1962 soundings. The validity of the data shown for the near midnight sounding on 5 May above the mesopause is still under investigation as our analysis is yet very preliminary.

Another noteworthy feature in both 5 May soundings is the remarkably high temperature in the stratosphere near 42 km. This fact is also brought out by the meteorological dropsonde observations on this day (figure 11). Since these high temperatures are again accompanied by meridional circulation (from the north) this case is very reminiscent of the stratospheric warming observed in January 1958 at Churchill.[2,4]

At this point it is of value to consider the wind and temperature structure together. In figures 1 and 7 the winds and temperatures are shown for all "winter" soundings, whereby "winter" is defined as that period where the winds are predominantly from the West. While the temperatures throughout this period are rather high and fluctuating in the *mesosphere*, the wind pattern throughout is uniformly zonal and unperturbed. During the transition period, however, where the high *stratospheric* temperatures are observed, a change in the wind pattern from zonal westerlies to meridional northerlies goes hand in hand with the high temperatures. This suggests that the causes of the sudden stratospheric warmings which seem to be connected very strongly with changes in the circulation from zonal to meridional are different from the causes which heat the mesosphere during most of the period of the uniform westerlies.

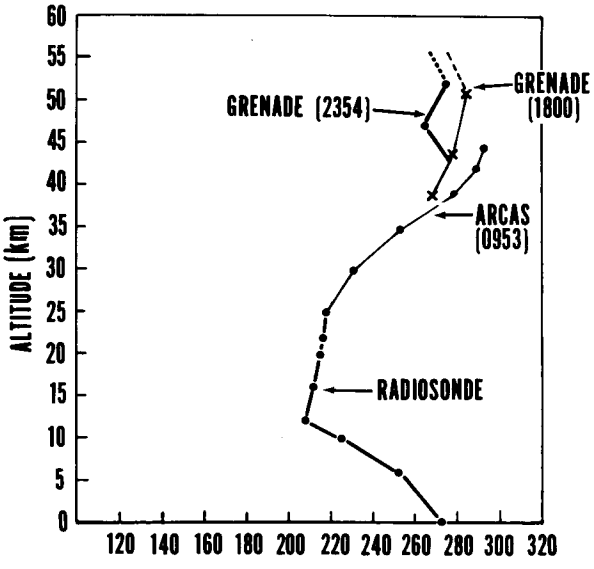


FIGURE 11.—Comparison of temperatures at Wallops Island (38°N) measured by Arcas and grenade technique, May 1961.

namely, the strong, but highly variable winds sandwiched between zones of relative calm resulting in extreme wind shears. Thus far, every sounding conducted has shown these wind shears between 90 and 110 km. Above 120 km greater uniformity seems to return but samples at these altitudes are too few to derive any definite circulation patterns (figure 5).

In figures 7 through 9 the individual temperature soundings are presented in seasonal groups. The winter atmosphere over Wallops Island is seen to be colder than the summer atmosphere below about 60 km. However, the warmer stratopause in summer connects to a very steep mesodecline resulting in a layer between 65 and 85 km which is colder in summer than in winter. This effect with appreciably greater magnitude was also observed at Churchill, Canada (59°N) by the IGY soundings.[2] Again, as in the Churchill observations, the mesospheric temperature structure exhibits a great deal more variability in winter than in summer and multiple temperature peaks between 50 and 90 km occur in every winter soundings, (figure 7). Because the temperature minimum at the tropopause is very shallow in winter and at the mesopause it is almost nonexistent, and because the temperature maxima at the ground and at the stratopause are much lower, the whole winter atmosphere up to 90 km approaches a more nearly isothermal average than the summer atmosphere.

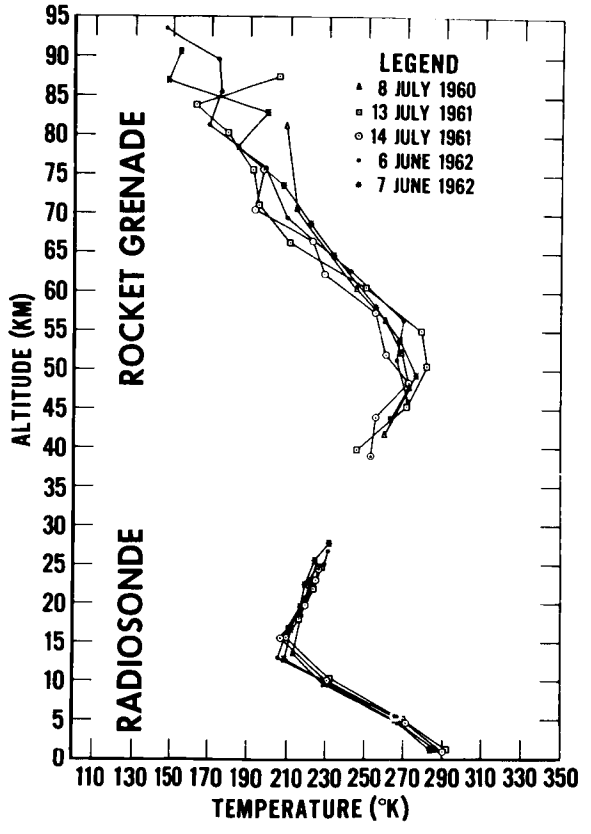


FIGURE 8.— 1960, 1961, 1962 summer temperatures up to 95 km at Wallops Island, Va. (38°N).

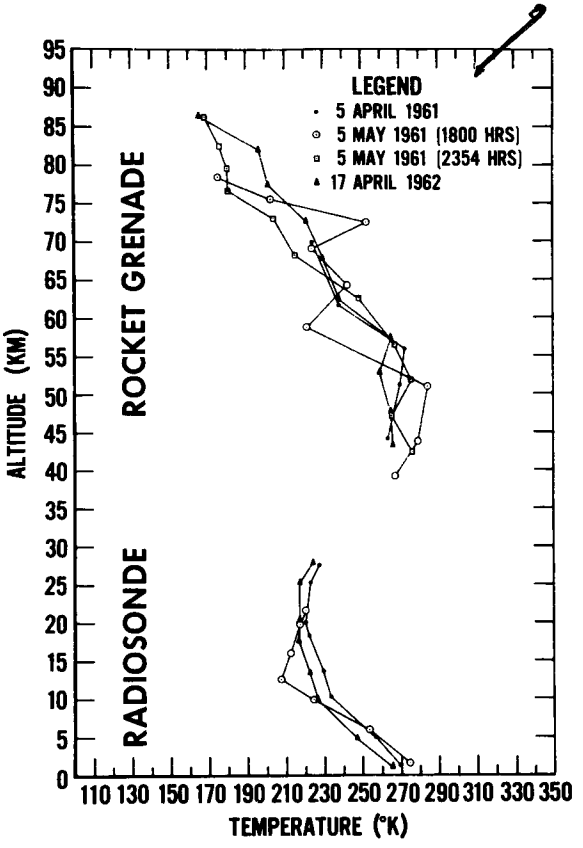


FIGURE 9.—Winter to summer transition period temperatures up to 95 km at Wallops Island, (38°N), April 1962, May 1961.

While large scale vertical temperature gradients are of much lesser magnitude in winter than in summer extremely large gradients on a smaller scale (5–10 km) can be found in winter. In fact, the smoothness and steepness of the summer lapse rate between 50 and 80 km is striking when compared to the perturbed but on the average very weak lapse rate in the same region in winter.

The temperature and wind profiles shown in figure 4 are the results of the only successful sounding during the summer to winter transition period. Although this small data sample constitutes very poor statistics, one may infer, by referring once again to the Churchill IGY results,[2] that the rather turbulent transition observed both at Wallops Island (5 May 1961) and a Churchill (27–29 January 1958) in spring cannot be seen in autumn. Figure 4 shows that the winds are extremely weak and that the temperature at 50 km lies near 260°K, typical for a wintertime stratopause, but without the typical wintertime temperature excursions in the mesosphere.

An interesting change in both the temperature and wind structure occurs in the 18 hour period between the two soundings on 13 and 14 July 1961. Although both soundings show typical summertime situations, the temperature at 50 km dropped by nearly 20°K during this period. The temperature decrease is accompanied by a change in the wind vector which exhibits the introduction of an appreciable southerly component on 14 July. Thus the temperature decrease may very well be attributed to advection of colder air from latitudes south of Wallops Island, a phenomenon previously observed at other latitudes.[2,15] Above 60 km temperatures from the two soundings agree well.

CONCLUSION

In conclusion, then, one finds that the atmosphere over Wallops Island (38°N) exhibits very similar features as those observed at Churchill (59°N) during IGY. Previously we had attributed these features to a "high latitude" atmosphere.[1] There are indications that the temperature differences between summer and winter at 50 km and 80 km which were in the average 17°K and 80°K respectively at Churchill are appreciably less at Wallops Island, (10°K and 25°K respectively. Nevertheless at both sites we find a warm winter mesosphere with a highly variable temperature structure, a feature not revealed in soundings at latitudes lower than Wallops Island. A striking result can be derived from all soundings conducted so far at any latitude. There seems to be a layer between 60 and 65 km where the temperature at all points at all seasons remains constant at about 235–240°K; this is surprising since immediately below and above, large temperature excursions take place both with latitude and season.

At this time one cannot reach a definite conclusion whether the differences found between the Churchill and Wallops Island soundings reflect a true latitudinal behaviour or whether they are due to the differences in time between these soundings; the Churchill soundings were performed during IGY at the height of the sunspot cycle while the Wallops Island results were obtained 4 to 5 years later. Simultaneous launchings at both locations now in progress will provide more information on this matter, although the assumption of a true latitudinal variation seems to be more likely.

REFERENCES

1. Nordberg, W., and Stroud, W. G., "Seasonal, Latitudinal and Diurnal Variations in the Upper Atmosphere," NASA Technical Note D-703, April 1961.
2. Stroud, W. G., Nordberg, W., Bandeen, W. R., et al., "Rocket-Grenade Measurements of Temperatures and Winds in the Mesosphere over Churchill, Canada," *J. Geophys. Res.* **65**(18): 2307–2323, August 1960.
3. Nordberg, W., and Stroud, W. G., "Results of IGY Rocket-Grenade Experiments to Measure Temperatures and Winds above the Islands of Guam," *J. Geophys. Res.* **66**(2): 455–464, February 1961.
4. Jones, L. M., Peterson, J. W., et al., "Upper-Air Density and an Abrupt Warming in the Mesosphere," *J. Geophys. Res.* **64**(12): 2331–2340, December 1959.
5. Ainsworth, J. E., Fox, D. F., and LaGow, H. E., "Upper-Atmosphere Structure Measurement Made with the Pitot-Static Tube," *J. Geophys. Res.* 66(10): 3191–3212, October 1961.
6. Spencer, N. W., "Density and Temperature in the Mesosphere above Churchill, Canada," (Submitted to J. Geophys. Res.).
7. Groves, G. W., and others, "Wind and Temperature Results Obtained in Skylark Experiments," in: Space Research: Proc. 1st Internat. Space Sci. Sympos., Nice, January 1960, Amsterdam: North-Holland Publ. Co., 1960, pp. 144–153.
8. Smith, L. B., "The Measurement of Winds Between 100,000 and 300,000 ft by use of Chaff Rockets," *J. Meterol.* **17**(3): 296–310, June 1961.
9. Kellogg, W. W., "Chemical Heating above the Polar Mesopause in Winter," *J. Meterol.* **18**(3): 373–381, June 1961.

10. HAURWITZ, B., "Frictional Effects and the Meridional Circulation in the Mesosphere," *J. Geophys. Res.* **66**(8): 2381–2391, August 1961.
11. MAEDA, K., "On the Heating of the Polar Upper Atmosphere," NASA Technical Report R-141, 1962.
12. AUFN Kampe, H., "Review of U. S. Meteorological Rocket Network Activities and Results," in: 1st Internat. Symps. on Rocket and Satellite Meterology, Washington, April 1962, Amsterdam: North-Holland Publ. Co. (To be published).
13. Data Reports, Meteorological Rocket Network, published by: The Meteorological Working Group Inter-Range Instrumentation Group (IRIG-MWG) U. S. Army Electronics Research & Development Activity, White Sands Missile Range, New Mexico.
14. MANRING, E. R., "Study of Winds, Diffusion, and Expansion of Gases in the Upper Atmosphere," Quarterly Progress Report, 1962, Submitted by Geophysics Corporation of America Under NASA/GSFC Contract No. NASW-396.
15. STROUD, W. G., NORDBERG, W., and WALSH, J. R., "Atmospheric Temperatures and Winds between 30 and 80 km," *J. Geophys. Research*, **61** (1956) 45.

SOME ASPECTS OF STRATOSPHERIC CIRCULATION DERIVED FROM METEOROLOGICAL ROCKET FIRINGS OVER THE UNITED STATES DURING THE WINTER OF 1961

MOHAMMAD RAHMATULLAH*

Goddard Space Flight Center

Data from six rocket launching sites in North America obtained from a series of meteorological rocket firings conducted between January 16 and February 23, 1961, were analyzed. Time cross-sections of winds at Wallops Island, Virginia, and Cape Canaveral, Florida, are presented and discussed in the light of synoptically significant situations as determined by the 10 mb chart. Temperature soundings in the lower stratosphere relevant to the typical situations are also considered. TIROS II (1960π1) radiation data for orbits 827 and 957 have been analyzed in order that the effect of ozone on stratospheric disturbances might be studied. It was found that the penetration of the mesospheric westerly jet to lower levels was associated with the formation of a marked trough of low pressure over the eastern United States and coincided with the occurrence of a severe cold period. No final stratospheric warming was observed and by mid-February the westerlies were back to full strength. A microstructure of the mesodecline at 63 km was observed at Wallops Island and at White Sands, New Mexico. The analysis of TIROS II radiation data pointed out that the effect of changes in the ozone layer is completely masked by the varying cloud pattern.

INTRODUCTION

This paper is primarily concerned with the study of stratospheric circulation over North America in the light of wind and temperature measurements from a series of meteorological rocket firings conducted between January 16 and February 23, 1961. The rocket results were reported by the Interrange Instrumentation Group (IRIG) in their quarterly report for the winter firings of 1961 and were prepared by the U.S. Army Signal Missile Support Agency, White Sands, New Mexico.[1] Data from six meteorological rocket launching sites in North America were studied: Cape Canaveral, Florida; Fort Churchill, Manitoba, Canada; Point Mugu, California; Tonapah Range, Nevada; Wallops Island, Virginia; and White Sands, New Mexico. The rockets fired include Arcas-Loki I and II and Hasp systems. The payloads were Robin spheres, parachute sondes, and Chaff. These rocket systems are supposed to provide wind measurements from 15 to 75 km, and the parachute sonde system, in addition, takes temperature measurements.

The operational problems involved in firing rockets, such as weather conditions, range safety, etc., make a synoptic firing schedule difficult. Therefore, the time interval chosen has to be based on the coincidence of the following factors: (1) the occurrence of a synoptically significant situation, as determined by the 10 mb charts; and (2) the availability of rocket data and other relevant information such as the radiation data from TIROS II (1960 π1). The primary objective in the present case was to study the polar vortex as related to the changes in temperature and wind pattern recorded by the various rocket firing ranges. It was also conceived that both the TIROS II radiation data collected during this

* This paper was written while Mr. Rahmatullah, regularly of the Pakistan Meteorological Department, was in residence at Goddard Space Flight Center under an international cooperation program.

period, and also temperature and wind data from a rocket grenade experiment fired on February 16, 1961, at Wallops Island might contribute to the analysis.

UPPER AIR CIRCULATION OVER NORTH AMERICA DURING JANUARY AND FEBRUARY OF 1961

The wind circulation in the stratosphere is generally governed by the polar-night vortex. The breakdown of the polar vortex is accompanied by some remarkable changes, of which the most important is perhaps the explosive stratospheric warming, first observed by Scherhag in 1952.[2] Rocket grenade measurements of temperature and winds over Fort Churchill, Canada, in January 1958 have shown that stratospheric

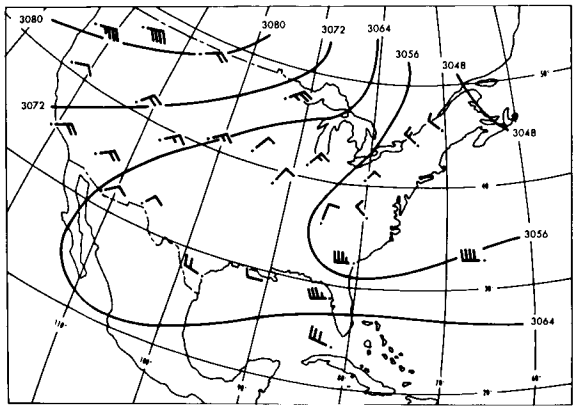


FIGURE 1.—Circulation at the 10 mb level, January 17, 1961.

circulation breakdown and associated warming were preceded by a breakdown of circulation throughout the stratosphere and lower mesosphere.[3] Teweles and Finger have also studied this phenomenon thoroughly for January 1958 and have found that the temperature south of Greenland rose from −70° to −39°C in a period of 4 days and that this rise was associated with major changes in the upper air circulation.[4] Generally, it has been observed that the polar vortex is not fully restored after a breakdown in the stratospheric circulation.

During January 1961 the upper air circulation at 10 mb over North America was greatly affected by the eastward movement of an anticyclone from

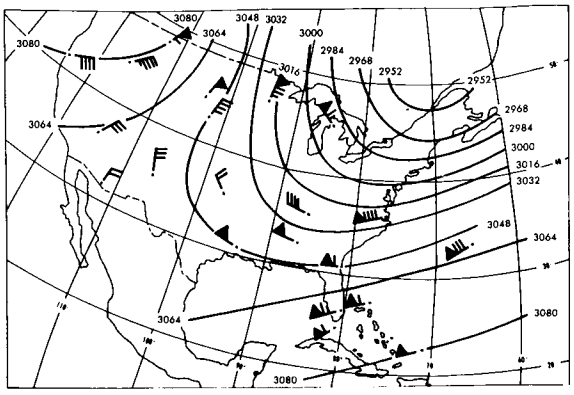


FIGURE 2.—Circulation at the 10 mb level, January 26, 1961.

the Alaskan region to the Hudson Bay area on January 9. The intensity of this antiyclone was weak compared to that of one in the previous year in which the height of the 10 mb surface over the Bering Strait reached values as high as 3192 geopotential decameters between January 7 and 11.[5] The anticyclone of 1961 retreated to the Aleutian region and it finally disappeared by the end of January. Circulation at the 10 mb level over most of Canada and the United States on January 17, 1961, was dominated by the retreating anticyclone (figure 1).

The weakening of the Aleutian high resulted in the migration of the cold vortex from Siberia to the Canadian sector of the Arctic during the last 10 days of January 1961. Initially a trough of low pressure developed which extended from Greenland to Labrador; later it formed into a separate

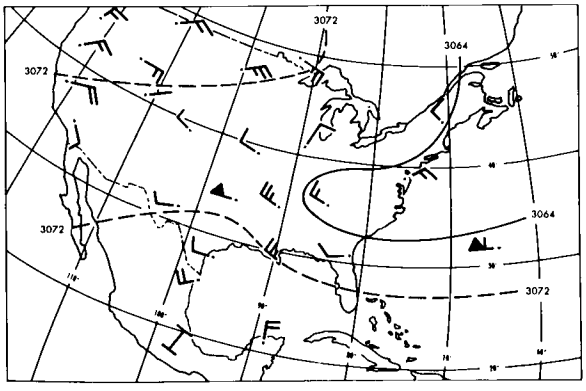


FIGURE 3.—Circulation at the 10 mb level, February 16, 1961.

cyclonic cell. Labitzke-Behr et al. pointed out that this cell, contrary to the main vortex, showed a warm core with temperatures up to −40°C, reached on January 25.[5] The synoptic pattern at the 10 mb surface for January 26, 1961, is illustrated in figure 2. On this date the eastern half of the United States was clearly under the influence of the cyclonic vortex. Circulation over the Pacific Coast and the adjoining Rockies was anticyclonic. By the end of January the wind circulation over the United States at the 10 mb level had become strongly zonal. Gradually the zonal flow weakened and by the first week of February a weak shear which had been over the Great Lakes region was moving eastward over the Atlantic seaboard. Then the anticyclonic circulation again developed over most of the western United States under the influence of a new Aleutian high which had been established on February 12. A typical example, illustrating this regime for February 16, is given in figure 3.

TIME CROSS-SECTION OF WIND AT WALLOPS ISLAND, VIRGINIA

Wallops Island is the only rocket firing range representative of the middle latitudes in the eastern United States. Rocket wind data from Fort Churchill, Canada, were available for 1 day in January and 6 days in February 1961. Since most of the latter firings did not exceed the height of 30 km, they were of limited use in this study. The time cross-section of zonal winds over Wallops Island is given in figure 4. It is obvious that the wind structure as revealed by this time cross-section is quite different from the mean zonal cross-section as computed by Murgatroyd.[6] Recently Batten, in the light of the discovery of the phenomenon of explosive warming observed in the months of January and February, tried to compute separate mean cross-sections for November to December and January to February. He also incorporated all available rocket data in his computation. The winter cross-section for Wallops Island for 1961 shows fair agreement with Batten's zonal cross section in that the tropospheric westerlies are bounded by a region of weak easterlies or zones of calm wind. However, two points stand out strikingly in contrast to Batten's model of mean zonal flow. The first is the penetration of the westerly maximum in the stratosphere to a level of about 35 km on or about January 25, 1961, and the second is the predominance of strong easterly flow during the last days of January. The highest velocity of the easterly wind recorded for the last days of January was 100 knots near 45 km on January 31, 1961. During

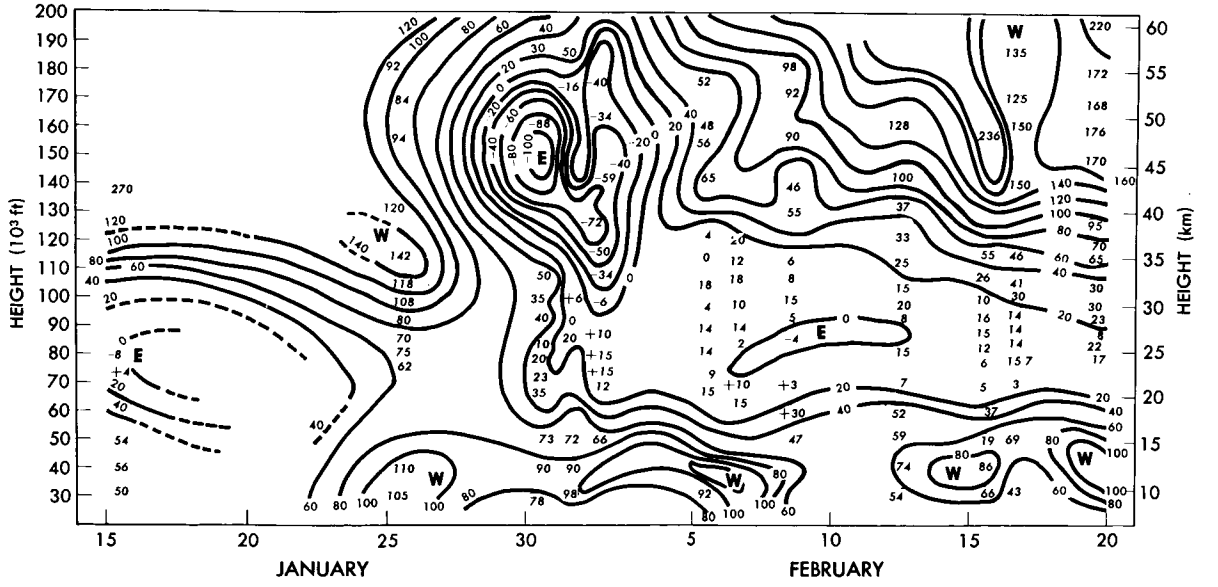


FIGURE 4.—Time cross section of zonal winds (knots), Wallops Island.

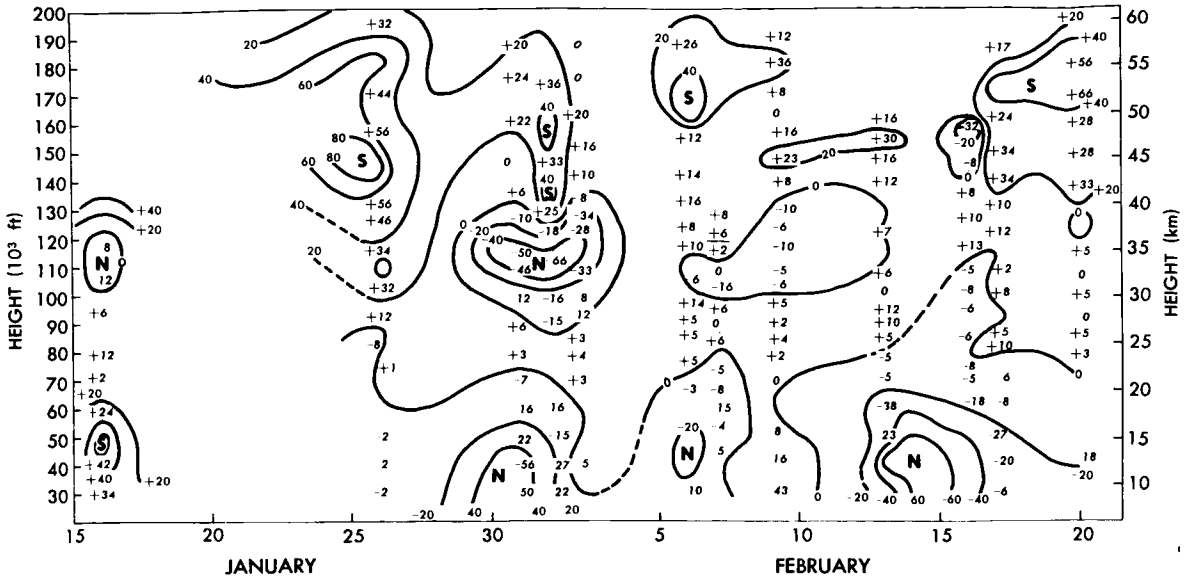


FIGURE 5.—Time cross section of meridional winds (knots), Wallops Island.

February the flow, though variable from day to day, generally corresponded to the standard winter pattern in which the westerlies are strong and increase with height. Apparently in this case, contrary to the previously observed breakdowns of the polar vortex, the polar westerlies returned to full strength. It is noteworthy that during the period of study no explosive warming in the stratosphere was observed. The penetration of strong westerlies to lower altitudes, near the stratosphere, about January 25, was associated with the presence of a deep trough over the eastern United States and Labrador at the 10 mb level. It is interesting that the severe cold and blizzardous period observed in the winter of 1961 coincided with the penetration of strong westerlies to the low stratospheric levels.

In considering the zonal maximum of the easterlies at about 45 km on January 31, 1961, it appears that the development of the warm core cyclonic cell near Labrador may have been influential in changing the flow pattern, since the presence of warm air to the north of Wallops Island would have given rise to thermal wind effects tending to cancel the westerly zonal flow. However, in the absence of temperature data over Wallops Island for that date, this interpretation may be taken as purely tentative.

The time cross-section for meridional winds at Wallops Island is given in figure 5. It indicates that southerly winds predominate in the upper stratosphere. The lower stratosphere is marked by weak variable winds. Meridional flow is not as strong as zonal flow. It can be inferred that the wind circulation in the upper stratosphere is predominantly a zonal flow over which the meridional component has been superimposed in association with upper air wave circulation.

TIME CROSS-SECTION OF WIND AT CAPE CANAVERAL, FLORIDA

The data from Cape Canaveral are limited to the period from January 16 to 25. The time cross-sections of zonal and meridional winds are given in figures 6 and 7, respectively. These figures show that strong westerly flow was well established at Cape Canaveral during the epoch of westerly wind penetration at Wallops Island. Westerly wind velocity as high as 100 knots was reported from Cape Canaveral on January 25 at about 33 km. The mean wind velocity in Batten's computed cross-section at the latitude of Cape Canaveral is only of the order of 15 to 20 knots. It demonstrates that the stratospheric disturbance is large and its influence on the wind pattern extends at least from 38° to 28°N. As in the case of Wallops Island the predominant meridional wind com-

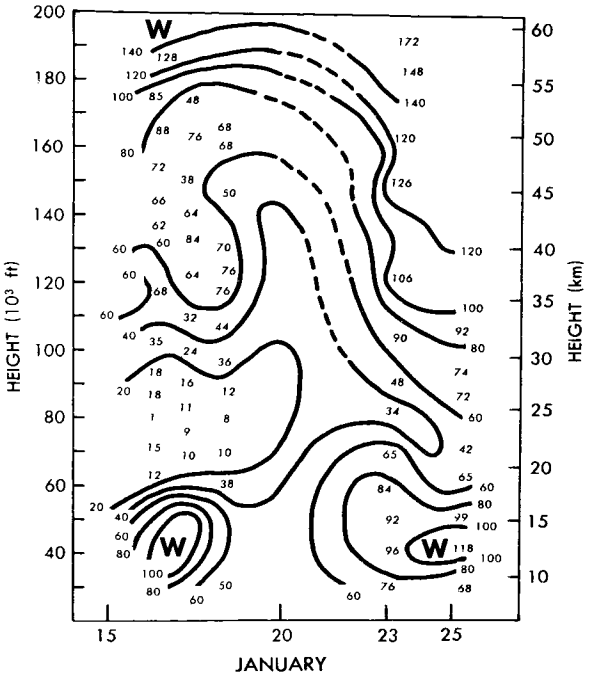


FIGURE 6.—Time cross section of zonal winds (knots), Cape Canaveral.

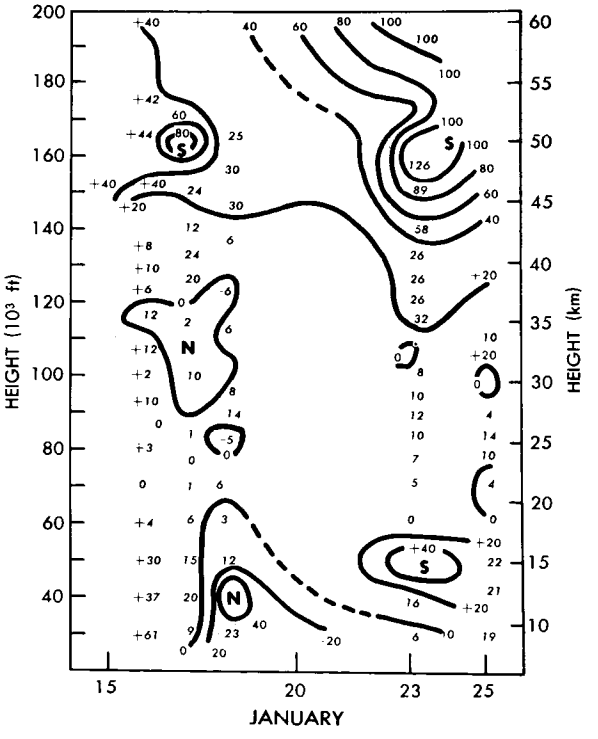


FIGURE 7.—Time cross section of meridional winds (knots), Cape Canaveral.

ponent in January is from the south. The lower stratosphere is marked by weak variable meridional wind components.

ZONAL WINDS AT WHITE SANDS, NEW MEXICO

The zonal wind pattern for White Sands, New Mexico, has been studied by Keegan,[8] who observed that in mid-January strong westerly winds were reported at approximately 60 km. The zonal westerlies decreased after January 18 and reached a minimum between January 20 and 23. By January 26 the zonal westerlies were reestablished at White Sands with 100 knot winds at about 36 km. Thus, the stratospheric disturbances which affected Wallops Island also caused the strengthening of westerly wind at White Sands. The influence of the stratospheric disturbances was only lightly felt at Point Mugu.

TEMPERATURE SOUNDINGS

Rocket temperature data for the period in consideration were available only from Cape Canaveral, Fort Churchill, Point Mugu, and, occasionally, White Sands. Temperature data from one rocket grenade firing on February 16 were available, from Wallops Island.[9] Therefore, it was difficult to find suitable dates for which data from more than two or three stations could be used to correspond to the synoptic situation as illustrated in figures 1–3. The close agreement between temperature data obtained from rocket firings and from radiosondes inspired confidence in analyzing at least one cross-section, for January 17–18, for which temperature data from Cape Canaveral, Point Mugu, and White Sands were available. For January 26 and February 16 rocket temperature data from two stations, as available, were plotted in order to compare the stations in regard to the temperature field in the stratosphere and lower mesosphere.

The temperature cross-section for January 17–18 is given in figure 8. The time of the rocket firing for Cape Canaveral was in the evening, whereas the firings for White Sands and Point Mugu were in the morning. In general, temperatures at White Sands were lower than at Cape Canaveral above about 36 km. These two ranges used the same sensors. There were no

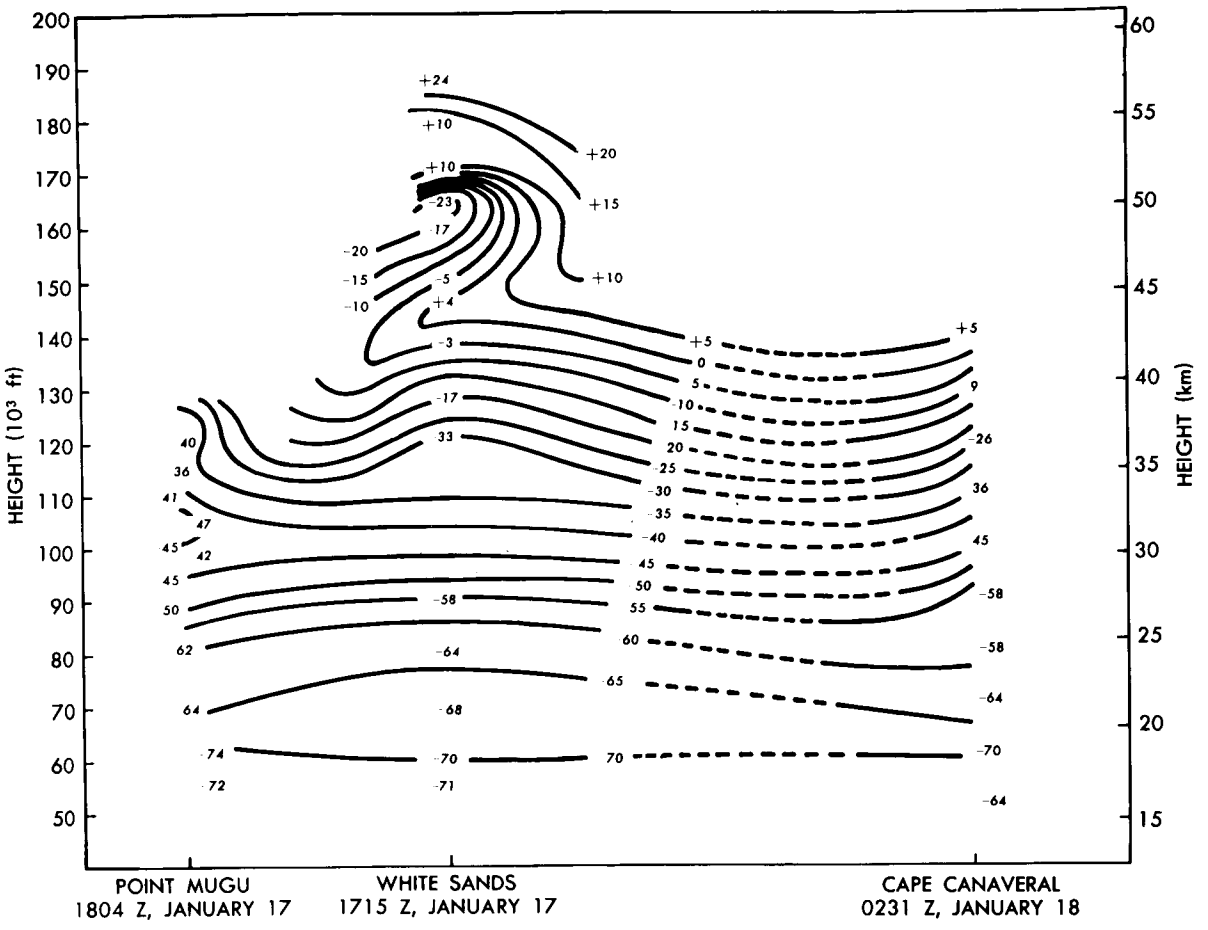


FIGURE 8.—Temperature cross section for Cape Canaveral, White Sands, and Point Mugu.

significant temperature differences between 15 and 36 km. The temperatures between 18 and 32 km were consistently higher at Point Mugu than at White Sands, possibly due to the fact that the two ranges used different sensors. The temperature recorded at 44 km, near the mesopeak, at White Sands was 4°C. Thereafter, the temperature decreased to about 50 km, when it started rising again. This peculiarity has been found in the winter soundings of higher latitudes. However, data should be interpreted with the understanding that these temperature sensors are very close to their limit of reliability at altitudes above 50 km.

Temperature data from rockets fired on January 26, 1961, are plotted in figure 9. Upper air temperature data from White Sands are not available for this date, but it can be assumed that

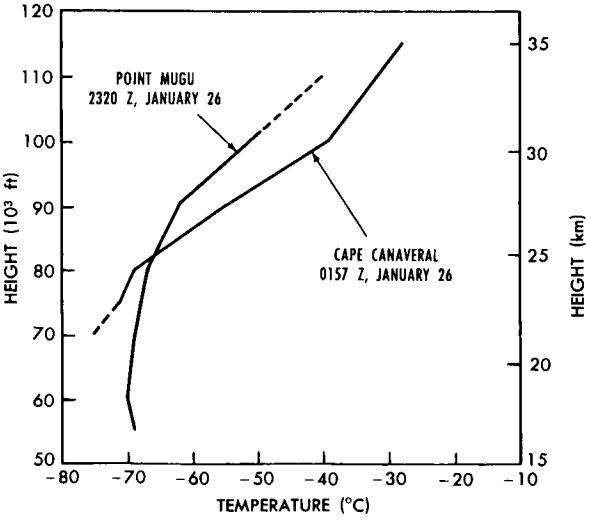


FIGURE 9.—Temperature soundings at Cape Canaveral and Point Mugu.

the temperature pattern there would not be substantially different from that at Point Mugu. It can be inferred that the disturbance which affected the United States in the fourth week of January caused appreciable temperature differences from east to west between 25 and 33.5 km in the lower latitudes.

The temperature soundings for White Sands and Wallops Island (rocket grenade data) for February 16 and 17 (figure 10) can be considered a example of a synoptic pattern with no strong cyclonic or anticyclonic patterns over the United States. On the 10 mb chart (figure 3) a weak trough of low pressure lies over the central United States and a weak anticyclonic pattern prevails over most of the western United States. The temperature data obtained from rocket firings agree so well with RAOB data that no doubt is left regarding their accuracy at the lower altitudes. The temperature in the stratosphere up to 30 km is generally about 5°C higher at Wallops Island than at White Sands. But at about 30 km this trend reverses. All of this can be explained on the basis of latitudinal differences between the two places and also is reflected in the mean wind cross-section.

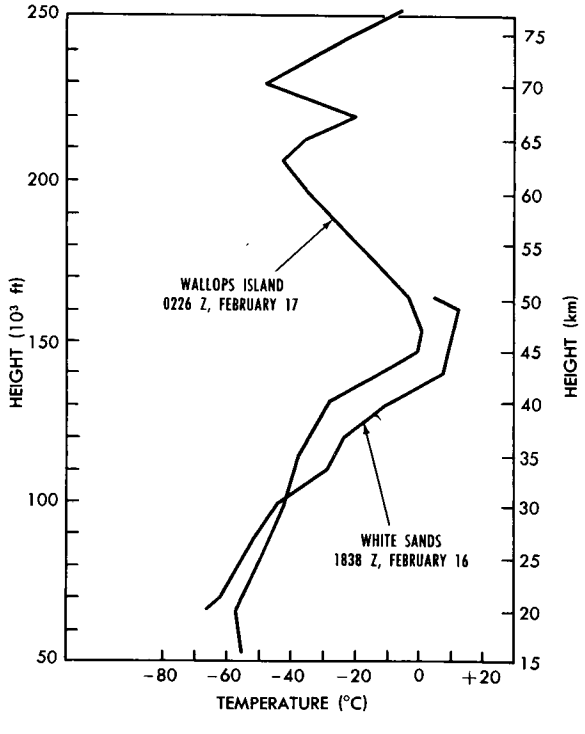


FIGURE 10.—Temperature soundings at Wallops Island and White Sands.

The mesopeak level at White Sands is reached at 46 km; the difference in the mesopeak temperature between White Sands and Wallops Island is 12°C. In the mesosphere the temperature sounding computed from the grenade firing at Wallops Island shows the same trend previously observed at very high latitudes.[3] At 63 km the temperature lapse rate reverses and a rise is experienced to 67 km. The temperature falls again after 67 km until the lowest value, −47°C, is reached at 70 km; it rises thereafter. It is difficult to evaluate the significance of this peculiar behavior of the wintertime temperature structure in the mesophere. Recently Nordberg published results of rocket grenade firings at Wallops Island which definitely confirm the existence of this secondary temperature maximum at these relatively low latitudes.[9]

RADIATION DATA FROM CHANNEL 2 OF TIROS II

In order to study the gross features of cloudiness and temperature conditions over the United States in the different regimes discussed earlier, radiation data obtained from channel 2 of TIROS II for orbits 827 and 957 were studied. Radiation registered by channel 2(8–12 μ) comes mostly from the ground or from cloud tops. The radiation data of January 18, 1961, obtained for orbit 827 are illustrated in figure 11. The analysis indicates very low values of radiant emittance, on the order of 14 to 16 watts/m^2, were measured over the north central United States. The thick black lines show the position of the surface front at 2100Z on January 18, which is fairly close to the time of the passage of the satellite over the United States. Scattered and broken cloud cover prevailed over the eastern states with surface temperatures from about 10°C to 15°C. But over much of the Middle West and Northwest the sky was practically overcast. Cloudiness decreased north of the front, from Wisconsin to the Dakotas, but surface temperatures were still well below freezing. It had been hoped that the TIROS II channel 2 data, which represent radiation received in the 8 to 12 micron region, might reflect

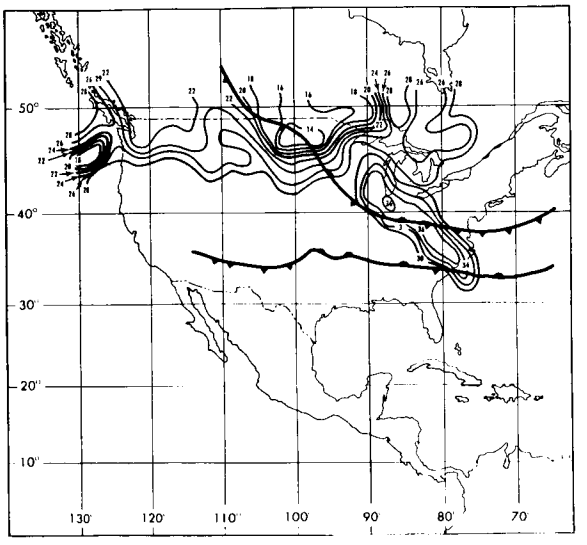


FIGURE 11.—Isolines of radiation (watts/m²), light black lines, for Tiros II orbit 827, 2110 *Z*, January 18, 1961, for channel 2 (8–12 μ). The heavy black lines indicate surface fronts at 2100 *Z*.

some variation in ozone content and distribution which would be related to the circulation phenomena described in the third, fourth, and fifth sections of this report. However, the synoptic situation described above causes such large variations in the temperatures of the emitting levels in the troposphere and at the earth's surface that any changes in the radiative transfer in the stratosphere would be completely masked by these lower level variations. Thus in this study the channel 2 data can be used only to describe the synoptic situation in the troposphere.

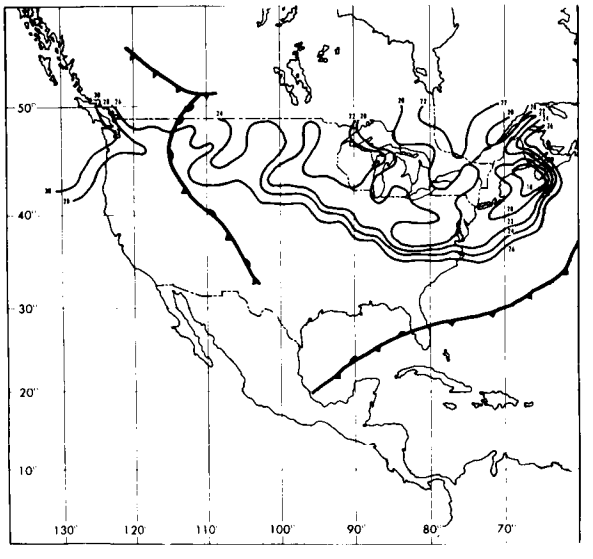


FIGURE 12.—Isolines of radiation (watts/m²), light black lines, for Tiros II orbit 957, 1755 *Z*, January 27, 1961, for channel 2 (8–12 μ). The heavy black lines indicate surface fronts at 1800 *Z*.

The same conclusion holds for orbit 957 on January 27, 1961, although in this case the radiation pattern is much more uniform (figure 12). Radiant emittances of 20 watts/m² cover practically all of the area from the Great Lakes to New England. The sky was overcast over the eastern states at 1800Z, the closest synoptic hour to the TIROS pass time. Over New England, cloud cover was broken to scattered, with temperatures ranging near −7°C. Around the Great Lakes, cloud cover was scattered to broken with increasing cloudiness toward the Northwest. Surface temperatures varied from −20° to −25°C over most of the Middle West and Northwest. This is the period characterized by the trough on the 10 mb chart over the eastern United States, but it must be concluded that the low radiation values scanned by channel 2 are indicative of the cloudiness and the bitter cold spell which gripped most of the eastern United States in the fourth week of January. It would have been desirable, of course, to study further the day to day variation in the radiation pattern, but it was not possible in this case because of the change in the orbital position of the satellite.

CONCLUSION

The study of the wind and temperature data from rocket firings during the winter of 1961 brings out the following important facts:

1. The upper stratosphere was subject to wind circulation, reflected on the 10 mb chart, which influenced the upper air circulation over extensive areas in both the north-south and east-west directions. The penetration of the mesospheric westerly jet to lower levels was associated with the formation of a marked trough of low pressure over the eastern United States. This penetration coincided with the occurrence of a severe cold and blizzardous period in that area during the fourth week of January.

2. The weakening of the polar vortex and the associated 10 mb trough ushered in a strong easterly flow which was later replaced by strong zonal westerlies. No final stratospheric warming was observed during the period of study and the westerlies had returned to full strength by the middle of February. This is contrary to the case where stratospheric warming had previously been observed in association with a breakdown of the polar vortex and the westerlies were not fully restored.[4]

3. Circulation changes in the stratosphere and the lower mesosphere affect the ozone content of the atmosphere. This phenomenon has been studied in detail by Wexler.[10] It was thought that the changes in the ozone layer might be reflected in the radiation data obtained for channel 2 of TIROS II for the typical regimes under study. But the radiation data were so greatly influenced by cloud cover that the effects due to changes in the ozone layer were completely masked by it.

ACKNOWLEDGMENTS

The author expresses his thanks to Dr. William Nordberg of Goddard Space Flight Center for assistance and guidance in the preparation of this paper and for making available all the data needed in this study. He is also thankful to Prof. Scherhag for the permission to reproduce figures 1–3 from the northern hemispherical charts published by the Free University of Berlin. He is indebted to Prof. Abdus Salam, Chairman, Pakistan Space Committee, and Dr. I. H. Usmani, Chairman, Pakistan Atomic Energy Commission, for sponsoring his visit to the United States.

REFERENCES

1. "Data Report of the Meteorological Rocket Network Winter 1961 Firings," U. S. Army Signal Missile Support Agency, White Sands Missile Range, IRIG-MWG No. 6–61, April 1961.
2. Scherhag, R., "Die Explosionsartigen Stratosphärenerwärmungen des Spätswinters 1951/1952," *Berichte des Deutschen Wetterdienstes in der U. S. Zone* No. 38: 51–63, 1952.
3. Stroud, W. G., Nordberg, W., et al., "Rocket-Grenade Measurements of Temperatures and Winds in the Mesosphere over Churchill, Canada," *J. Geophys. Res.* **65**(8): 2307–2323, August 1960.
4. Teweles, S., and Finger, F. G., "An Abrupt Change in Stratospheric Circulation Beginning in Mid-January 1958," *Monthly Weather Rev.* **86**(1): 23–29, January 1958.
5. Labitzke-Behr, K., Petzoldt, K., and Scherhag, R., "Preliminary Daily Northern Hemisphere 10-Millibar Synoptic Weather Maps of the Year 1961. Part 1: January–March," *Meteorologische Abhandlungen des Instituts für Meteorologie und Geophysik der Freien Universität Berlin* **20**(1): 1–99, 1961.
6. Murgatroyd, R. J., "Winds and Temperatures Between 20 km and 100 km—a Review," *Quart. J. Roy. Meteorological Soc.* **83**(358): 417–458, October 1957.
7. Batten, E. S., "Wind Systems in the Mesosphere and the Lower Ionosphere," Rand Corp. Paper P-2018, June 13, 1960; also *J. Meteorology* **18**(3): 283–291, June 1961.
8. Keegan, T. J., "Large Scale Disturbances of Atmospheric Circulation Between 30 and 70 Kilometers in Winter," *J. Geophys. Res.* **67**(5): 1831–1838, May 1962.
9. Nordberg, W., and Smith, W. S., "Preliminary Measurements of Temperatures and Winds above 50 km over Wallops Island, Virginia," NASA Technical Note D-1694 (in press); this report will also be published under the title "Grenade and Sodium Rocket Experiments at Wallops Island, Virginia," in: *Proc. Internat. Sympos. on Rockets and Satellite Meteorology, Washington, April* 1962, Amsterdam: North Holland Publ. Co. (in press).
10. Wexler, H., "A Meteorologist Looks at the Upper Atmosphere," in: *Atmospheric Explorations: Papers of the Benjamin Franklin Memorial Symposium of the American Academy of Arts and Sciences*, ed. by H. G. Houghton, Cambridge: Technology Press of Massachusetts Institute of Technology, 1958, pp. 79–100.

STRUCTURE OF PLANETARY ATMOSPHERES

S. I. RASOOL

Goddard Space Flight Center

NOMENCLATURE

- A albedo, fraction of solar radiation reflected by the planet; total integrated albedo corresponds to the albedo over the whole solar spectrum (ultraviolet, visible, and infrared), whereas visual albedo applies only to the visible radiation
- C root mean square velocity of escaping particles
- C_p specific heat of a gas at constant pressure
- G universal gravitational constant
- g acceleration due to gravity
- H $k\bar{T}/\bar{m}g$, scale height
- k Boltzmann constant
- M mass of the planet
- m mass per particle
- mb 10^{-3} bar $= 10^3$ dyne/cm^2
- N_e electron density, electrons/cm^3
- n_c number density of the atmosphere which would exist at the surface if the atmosphere at all heights were isothermal at the temperature of the exosphere
- n_0 number density of the atmosphere at the surface, mole/cm^3
- R radius of the planet
- r distance of the planet from the sun
- $\bar{T}$ mean temperature, °K
- T_0 temperature at reference level, °K
- T_c temperature at the level of escape or exosphere, °K
- T_e effective blackbody temperature of a planet $\propto r^{-1/2}\,(1-A)^{1/4}$
- T_G temperature of the surface of planet
- T_z temperature at level z
- t_e time in which the atmospheric density of a planetary atmosphere will fall to $1/e \sim 1/2.7$ of its original value due to the gravitational escape of gases
- z altitude, km
- β adiabatic temperature gradient
- κ mass absorption coefficient, cm^2/g
- ρ density of the atmosphere at any given level, g/cm^3
- ρ_0 density of the atmosphere at reference level, g/cm^3
- τ_0 total optical thickness of the atmosphere $= \int_0^\infty \kappa\rho\, dz$
- τ_z optical thickness of the atmosphere above the level $z = \int_z^\infty \kappa\rho\, dz$

INTRODUCTION

Intensified research activity in the field of planetary atmospheres during the last few years has led to a number of new results that have forced a revision of our previous understanding of the structure of the atmospheres of Venus and Jupiter.

Recent measurements of the intensity of the radiation emitted by the planet Venus in the centimeter wavelength show that it corresponds to thermal radiation of temperature of ~600°K.[1] Since radiation in the decimeter region probably passes unattenuated through the atmosphere and clouds of Venus, it generally is assumed that the measured temperature refers to the surface of the planet. Also, there is evidence that the atmosphere of Venus may not be composed predominantly of CO_2, as previously estimated,[2] but is present only as a minor constituent in a mixing ratio of 5%. The interpretation of these measurements has turned out to be one of the most interesting problems in planetary science at the present time.

It has been shown that Jupiter is the source of yet another type of intense radiation recently observed in the decimetric wavelength region.[1]

There is also an indication that the Jovian atmosphere may be predominantly helium and not hydrogen, as so far has been believed.[3] In the case of Mars, renewed interest, both theoretical and observational, has led to revised atmospheric models.[4–6] The purpose of this article is to review the properties of the atmospheres of Mars, Venus, and Jupiter in the light of the most recent observational results.

General Considerations

Temperature, density, and composition are the three essential parameters that determine the structure of a planetary atmosphere. Of these, temperature is the most significant because it directly reflects the processes of energy absorption in the atmosphere and at the ground. The vertical temperature structure of the Earth's atmosphere is fairly well known,[7] and, therefore, before entering into a detailed discussion of the structure of the atmospheres of other planets, it perhaps will be desirable to understand the factors that determine Earth's observed temperature profile shown in Fig. 1.

Temperature

The solar radiation flux, with an effective blackbody temperature ~6000°K, reaching the top of Earth's atmosphere, has a value of $\sim 1.4 \times 10^6$ erg/cm^2/sec. Part of this radiation immediately is "reflected" back to space by clouds and the atmosphere and does not play any role in the energy balance of the planet. The fraction of the solar radiation thus reflected back to space is known as the albedo of the planet, and, in the case of Earth, it has been estimated to be 0.39 (e.g., see Ref. 8). Most of the remaining 61% of the solar radiation, composed mainly of the visible part of the spectrum, penetrates down to the ground and heats the surface to a certain temperature denoted as T_e or the effective temperature of the planet. In the case of fast rotating planets (e.g., Earth, Mars, Jupiter),

$$T_e^4 \propto (SC/4) \cdot (1-A)$$

where SC is the solar constant or the radiation flux received at the top of the atmosphere and A is the albedo of the planet over the whole solar spectrum. For the values of SC and A just given, the T_e for Earth is 245°K.

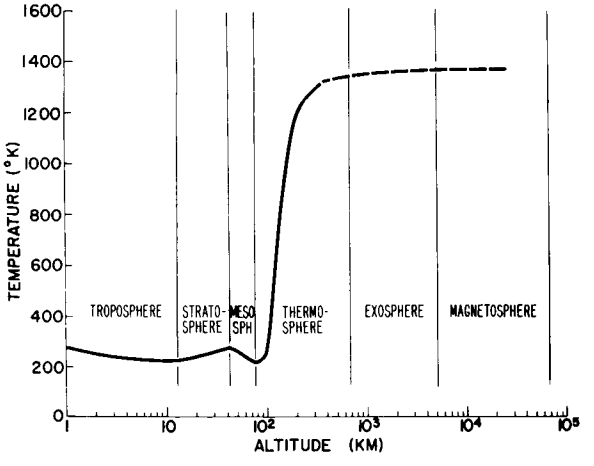


FIGURE 1.—Temperature profile of Earth's atmosphere (after Jastrow and Kyle, 1961[7]).

The surface radiating at this relatively low temperature emits primarily in the far infrared. A large fraction of the radiation emitted by the surface is absorbed immediately by molecules of CO_2 and H_2O present in the lower layers of Earth's atmosphere due to the excitation of intense vibration-rotation and pure rotation bands. A part of this absorbed energy in the infrared is returned to the surface and provides extensive heating of the ground, raising the surface temperature to the observed value of ~290°K.

This additional heating of the surface by the return of infrared from the atmosphere is referred to as the "greenhouse effect" and is an essential element in understanding the radiation budget of the planet in general and of the lower atmosphere in particular.

The lower atmosphere of Earth is in radiative and convective equilibrium, and the temperature decreases with height at a rate of ~6.5°K/km.

The adiabatic temperature gradient is given by $\beta_a = -g/C_p$, where C_p is the specific heat of air at constant pressure. For Earth's atmospheric composition, $\beta = -9.6$°K/km, but because of the eventual condensation of atmospheric water vapor in the form of clouds, the latent heat released in the atmosphere reduces the actual temperature gradient to the observed value of −6.5°K/km.

This region of negative temperature gradient extends up to an altitude of ~12 km and is

referred to as the *troposphere*. The top of the troposphere is denoted as the *tropopause*, and the temperature at this level is of the order of ~200°K. Above the tropopause, however, the atmosphere is mainly in radiative equilibrium. Because the water vapor precipitates out at the low temperature of the tropopause, there is a very small amount of water vapor in the atmosphere above to provide any significant infrared opacity. The temperature therefore remains in the neighborhood of ~200°K and then rises subsequently because of the direct absorption of the solar ultraviolet radiation by the traces of ozone present at ~50 km.

The atmospheric region between the tropopause and the level of the secondary temperature maximum (at 50 km) therefore has a positive temperature gradient and is stable against convection. It is referred to as the *stratosphere*.

Above the ozone layer, the temperature decreases again with altitude to a minimum value of ~180°K at ~80 km (*mesopause*).

Above 100 km, a region of strong heating sets in which results from the photodissociation of O_2 and the photoionization of N_2 and O by the solar radiation in the far ultraviolet. This is the region of the *thermosphere*. At the top of the thermosphere, the temperature approaches a constant value of ~1500°K.

The *exosphere* is defined as the region where the atmospheric density and the probability of collisions is so small that the particles execute ballistic trajectories in the gravitational field of the planet. The base of the exosphere is a level above which there is only one collision for a particle moving vertically away from the planet. In the case of Earth, the base of the exosphere is located at an altitude of ~700 km, where the density is of the order of ~10^6 particles/cm^2. The exospheric region extends up to the distance where the atmospheric density falls to the value of ~100 particles/cm^2, the average density of the interplanetary medium. On this criterion, the outer boundary of the exosphere lies near 5000 km. This can be considered as the limit of Earth's neutral atmosphere.

Charged particles trapped in Earth's magnetic field, however, populate a region beyond this, called the *magnetosphere*, which has been found to extend to several Earth radii and is known more commonly as Van Allen radiation belt.

The structure of Earth's atmosphere, as outlined in the foregoing, may serve as a guide for understanding and describing the atmospheres of other planets. The different circumstances under which other atmospheres may have evolved, however, no doubt will lead to substantially different structures with new regions and layers whose relationships bare only slight resemblance to Earth's atmosphere. Nevertheless, if one can understand the basic physics of our own atmosphere, it is possible to go quite far in inferring the elemental properties of the atmospheres of other planets, taking into account differences in composition, planetary mass and radius, and distance from the sun, whose radiation spectrum plays a very important role in atmosphere structure.

Composition

The condensation of planets from a gaseous mixture of solar composition initially would have given a planet an atmosphere composed predominantly of hydrogen and helium with relatively small amounts of CH_4, NH_3, H_2O, CO, O_2, N_2, CO_2, etc. Being the lightest of all elements, hydrogen and helium diffuse out to the fringes of the atmosphere and, depending on the size, mass, and temperature of the planet, eventually escape into the interplanetary medium. At any one time, therefore, the gross composition of a planetary atmosphere depends upon the steady state established as a result of the "escape velocity" of the planet and the mean thermal velocity of the atoms and molecules in the outer regions of the atmosphere. The atmospheric constituents are replenished by the exhalation of gases from the crust, and to some extent its composition is also modified by the solar ultraviolet radiation through the photodissociative and ionizing processes usually prevalent in the upper atmosphere. Factors like the chemical reactions of the gases in the atmosphere with the solid material of the crust, the capture of gaseous constituents from the interplanetary medium, and the radioactive processes in the interior of the planet also may contribute to the composition of the atmosphere.

From the considerations of Jean's theory of gravitational escape of gases, one can grossly estimate the present composition structure of Earth's atmosphere. According to Spitzer,[9] the time t_e in which the density of an atmospheric constituent of molecular or atomic weight m will fall to e^{-1} of its original value is given by

$$t_e = [B(6\pi)^{1/2}C/3g](e^y/y) \qquad (1)$$

where $B = n_0T_0/n_cT_c$ and $y = GmM/kT_cR$; n_0 and T_0 are the particle number density and temperature, respectively, at the ground level, whereas n_c and T_c are the same parameters at the escape level, which, as described before, is the base of the exosphere; R is the radius of the planet, and C is the root mean square velocity of the escaping particles.

For an exospheric temperature of ~1500°K, hydrogen would have escaped from Earth in $\sim 10^5$ yrs. As the age of the planets is $\sim 3\times 10^9$ yr, the absence of hydrogen in the atmosphere of Earth is comprehensible. Hydrogenated gases like CH_4 and NH_3, which are susceptible to dissociation by the solar ultraviolet radiation, also would lose their hydrogen, and the carbon and nitrogen thus liberated probably will remain in the atmosphere as CO_2 and N_2.[10]

The case of water is different. Because of the low temperature of the tropopause, almost all of the terrestrial water vapor remains confined to the troposphere. An extremely minute amount, corresponding to the saturated vapor pressure of 200°K, will traverse the tropopause and will will be susceptible to dissociation in the upper atmosphere. This special circumstance, according to Urey[10], preserves the water on our planet. Urey also has argued that the present amount of free oxygen can be accounted for by the amount of water vapor so far dissociated in the upper atmosphere.

In the case of He, the time of escape from Earth is of the order of 10^8 yr, which, because of the possible uncertainty in the assumed escape level temperature, cannot be interpreted as a definite indication of complete absence of He in Earth's atmosphere. In fact, a layer of helium at an altitude of ~1600 km recently has been detected by satellite investigations.[11]

From Eq. (1), it can be concluded that exospheric constituents of molecular or atomic weight >6 probably will be retained by Earth up to the present time. These considerations indicate that H_2 has escaped, and hence the present atmosphere of Earth should be oxidized state. Table 1 shows the observed composition of Earth's atmosphere.

TABLE 1.—*Sea Level Composition of Earth's Atmosphere*

Constituent	% volume	Molecular wt (O = 16.00)
N_2	78.09	28.016
O_2	20.95	32.000
A	0.93	39.944
$(H_2O)^a$	0 to 2.0	18.016
CO_2	0.03	44.010
Ne	1.8×10^{-3}	20.183
He	5.24×10^{-4}	4.003
Kn	1.0×10^{-4}	83.7
H_2	5×10^{-5}	2.016
N_2O	2.5×10^{-5}	44.032
CO	2.0×10^{-5}	28.010
Xe	8.0×10^{-6}	131.3
CH_4	5×10^{-6}	16.035
$(O_3)^b$	0 to 7×10^{-6}	48.000

[a] Above the troposphere, the amount of H_2O is always very small.

[b] O_3 reaches a maximum of 1 to 3×10^{-4}% between 20 and 30 km.

Nitrogen, oxygen, and argon make up more than 99.9% of Earth's atmosphere. Though the almost complete absence of hydrogen and helium perhaps can be explained by the gravitational escape of gases, the presence of free oxygen and the relative absence of CO_2 in the terrestrial atmosphere are problems related to the presence of life on Earth and the possible reactions of atmospheric gases with the crust. This subject very recently has been discussed comprehensively by Urey[10] and therefore will not be treated here in any detail.

The mean molecular weight for the atmospheric composition given in table 1 is ~29. Earth's atmosphere is supposed to be mixed up to an altitude of ~80 km, and the composition remains nearly constant except for H_2O and O_3, as indicated in the table.

The total pressure at the surface due to this atmosphere is $\sim 10^6$ dyne/cm^2 (1000 *mb*) and

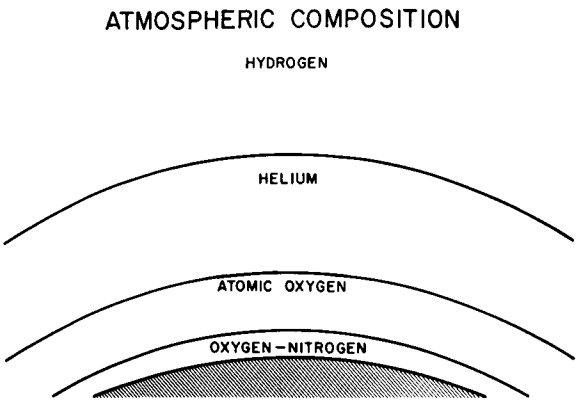


FIGURE 2.—Composition of the upper atmosphere of Earth (after Jastrow, 1962).

the number density at the ground $\sim 2 \times 10^{19}$ mole/cm^3. In a nearly isothermal atmosphere, the variation of density with height is given by the formula expressing hydrostatic balance:

$$\rho = \rho_0 \exp\{-[(h-h_0)/k\bar{T}]\bar{m}g\} \qquad (2)$$

in which ρ and ρ_0 are the densities at height h and h_0, $\bar{m}$ is the average molecular weight per particle, $\bar{T}$ the average temperature between h and h_0, g the acceleration of gravity, and k the Boltzmann constant.

The quantity $k\bar{T}/\bar{m}g$ is known as the scale height H of the atmosphere; at an altitude of one scale height, the density being reduced by a factor of e. For a mean temperature of 250°K in the first 80 km of the atmosphere, the atmospheric scale height is $\sim$8 km.

Above 120 km, diffusive separation sets in and the pressure of each constituent varies in accordance with a scale height calculated for its own molecular weight. At the boundary of this domain, at a height of 120 km, nitrogen and oxygen are the major constituents, with N_2 dominating the ratio 4:1, as on the surface. At higher altitudes the relative concentration of oxygen rapidly increases, and by 300 km the surface proportions are reversed, with oxygen now dominating in the ratio of 3:1.

The reason for this circumstance is that O_2 can be dissociated by absorption of solar ultraviolet, with relatively high probability, yielding atomic oxygen with half the weight per particle and, therefore, with twice the scale height and a correspondingly smaller rate of decrease of concentration with increasing altitude. Molecular nitrogen also undergoes photodissociation, but with a much smaller probability than for oxygen; hence it remains in primarily molecular form up to greater altitudes. Thus, again the twice-as-heavy nitrogen settles out of the atmosphere with relative rapidity. At 300 km, for example, less than 1% of the N_2 is in atomic form, whereas 99.5% of the oxygen is atomic.

For this reason, atomic oxygen becomes the principal atmospheric constituent above 200 km and continues to dominate the composition of the upper atmosphere up to about 1000 km. Above that level, the lighter gases, hydrogen and helium, emerge as the principal constituents because of their very large scale heights, although they are present only in trace amounts at lower altitudes. At the greatest altitudes, all gases except hydrogen have settled out of the atmosphere, and this lightest gas dominates, until finally the hydrogen atmosphere merges into the interplanetary medium at a distance of some 2 to 10 Earth radii. Figure 2 shows the variation of atmospheric composition with altitude.

Other Planets

It is interesting to use one's knowledge of the physical properties of Earth's atmosphere to interpret the observational results obtained for the other planets. In a detailed analysis of the problem, Urey[10] has surveyed the properties of the atmospheres of the other planets in the light of observational evidence available until 1958 and has arrived at many interesting conclusions regarding the origin of planetary atmospheres. Using improved techniques of optical and radio astronomy, many new observational results have been obtained in the past four years. In order to fit the observed data, new model atmospheres for Venus, Mars, and Jupiter have been derived. The properties of the atmospheres of these planets therefore will be discussed in more detail.

MARS

Composition and Pressure

Table 2 summarized reliable physical data for Mars. From the considerations of the cosmic

abundance of elements, the gravitational escape of gases, and the temperature environments of Mars, Kuiper[12] has given a list of gases which probably could be present in the atmosphere of Mars; they are given as follows in order of predicted relative abundance: CO_2+CO; (NO); $N_2(N_2O)$; (COS); (SO_2); H_2O; O_2+O_3.

These gases shown within parentheses may have been dissociated photochemically and hence their abundances reduced greatly. The rare gases like argon, xenon, and neon also would be present, but their abundances are very uncertain, especially that of argon, because it is generated continuously by radioactive decay of K_{40}. Keeping in mind these possible atmospheric constituents, the observational evidence of the composition of Martian atmosphere will now be discussed.

The only gas that has been detected spectroscopically up to now in the atmosphere of Mars is carbon dioxide.[12] Its abundance has been estimated by Goody and Grandjean[13] as 35 m-atm (where m-atm is the thickness of a homogeneous atmosphere in meters at 0°C and 760 mm). Thorough search for water vapor has failed to give a positive result, and recent spectroscopic observations by Kiess et al.[14] set an upper limit of 8×10^{-3} g/cm^2. From estimates of the thickness and dimensions of the polar caps, de Vaucouleurs[15] estimates the amount of water vapor at any time in the atmosphere of Mars to be $\sim10^3$ g/cm^2, which is below the amount observable from Earth, even from high altitude ballons. Kuiper[12] has looked for SO_2, O_3, N_2O, CH_4, and NH_3 but has failed to detect them, giving the following upper limits to the possible abundances of these gases

TABLE 2.—*Astronomical Data for Mars*

Mean distance from sun	1.52 a.u.
Mean equatorial diameter	6790 km
Length of day	1.0012 Earth day
Length of year	1.8808 Earth yr
Mass	0.1078 (Earth 1)
Mean density	3.90 g/cm^3
Gravity	377 cm/sec^2
Total integrated albedo	0.26±0.02 (Ref. 81)
Effective blackbody temperature	209°K

TABLE 3.—*Probable Composition of the Martian Atmosphere*

Gas	Amount, m-atm	% volume
N_2	~1675	95.0
A	~50	2.5
CO_2	~35	2.0
O	<2.4	<0.15
H_2O	$<2\times10^{-3}$g/cm^2	...

in the Martian atmosphere:

N_2O	2.0 m-atm
NH_3	0.2 m-atm
CH_4	0.1 m-atm
O_3	5×10^{-4} m-atm
SO_2	3×10^{-5} m-atm

Dunham[16] estimates the amount of O_2 on Mars to be less than 0.15% of the amount on Earth, which gives an upper limit of 2.4 m-atm. Recently Sinton[17] has studied the absorption spectra of Mars in the 3 μ region and has given an upper limit to the amount of NO_2 and N_2O_4 as 2.2 m-atm in the atmosphere of Mars.

De Vaucouleurs[15] observed the brightness of Martian features at various angles and, from the amount of light scattered, estimated the surface pressure of Mars to be 80±13 *mb*. Dollfus,[18] by polarization measurements, finds the surface pressure to be approximately 85 *mb*. The most likely value, according to Urey,[10] is 85±10 *mb* which amounts to an atmospheric mass per unit area of 230 g/cm^2, corresponding to a total atmosphere of 1760 m-atm.

If one assumes 35 m-atm of CO_2, then the remaining unidentified constituents of the atmosphere are probably mostly nitrogen with traces of argon. These gases lack absorption spectra in the observable part of the solar spectrum but are abundant cosmically.

The probable composition of the Martian atmosphere is shown in table 3. This atmosphere has a mean molecular weight of 28.5. With a surface pressure of 85 *mb* and an approximate temperature of 210°K, the surface density would

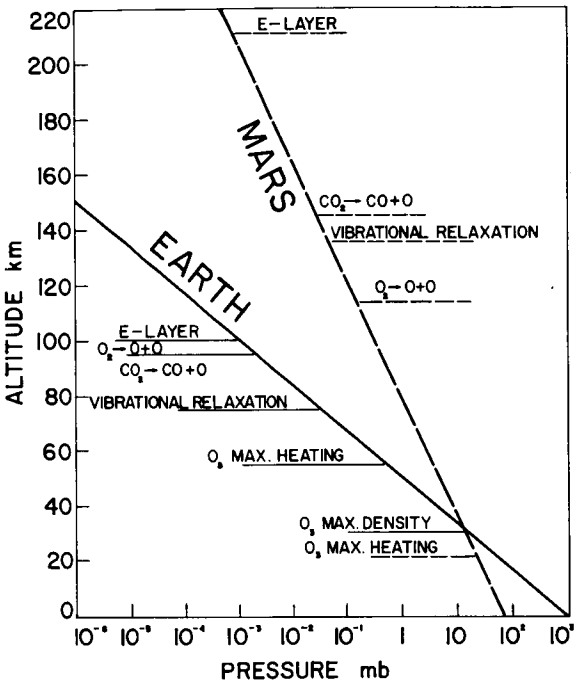


FIGURE 3.—Pressure vs. altitude for Mars and Earth (after Goody, 1957[19]).

be approximately 2×10^{18} particles/cm^3, which, for an isothermal atmosphere, should decrease exponentially with altitude with a scale height of ~20 km. The vertical distribution of pressure for this model has been calculated by Goody[19] and is shown in Fig. 3. Results of a similar computation for Earth, assuming an isothermal atmosphere at 250°K, also are shown in Fig. 3 and are in accord with recent rocket measurements. It is interesting to note that, although the pressure at the surface of Mars is about $\frac{1}{12}$ of that of Earth, at an altitude of 31 km the two atmospheres have the same pressure, and above this height the pressure in the Martian atmosphere is greater than the pressure at the corresponding height in Earth's atmosphere. Because the decrease of density with height in the atmosphere of Mars is almost $2\frac{1}{2}$ times slower than in Earth, the levels of the ionosphere and thermosphere on Mars would be much higher than on Earth. In Fig. 3, Goody has indicated the corresponding levels at which photochemical reactions take place.

Temperature

Planetary temperatures usually are estimated by measuring the infrared radiation emitted by the planet. A large part of the infrared spectrum, however, is absorbed by the water vapor and CO_2 present in Earth's atmosphere. Ground-based observations of the planets in the infrared are therefore confined in the 8 to 12 μ region where Earth's atmosphere is relatively transparent. This spectral region in the infrared is therefore known as the atmospheric "window."

Extensive temperature measurements of Mars by infrared radiometry in this "window" of Earth's atmosphere have been made since 1926, and the results to date are quite consistent. Since the Martian atmosphere also should be largely transparent in the 8 to 12 μ "window," except for the weak absorption by CO_2 at 9.4 and 10.4 μ, it is assumed that the radiation intensity measurements in this wavelength region refer to the surface of the planet. Adel[20] and recently Hess,[21] however, have pointed out that the presence of relatively large amounts of CO_2 in the Martian atmosphere makes all measurements of temperature on Mars doubtful, especially those near the limb of the planet. The argument is that, as the atmosphere probably will be cooler than the surface, emission from the CO_2 bands in the "window" will make the measured temperatures lower than the surface temperature.

From the known intensities of CO_2 bands at 9.4 and 10.4 μ,[22] one can estimate the contribution of CO_2 in the absence of water and other absorbing gases. Assuming a surface temperature of 300°K, the correction due to 35 m-atm of CO_2 at an average pressure of ~50 *mb* and at a mean temperature of 240°K is less than 3°K. One therefore can assume that, except for measurements made at limb, the temperatures measured in the 8 to 12 μ window refer to the surface of the planet.

De Vaucouleurs[23] has summarized all the available temperature measurements made until 1952, and they appear to be consistent within the possible observational error. Sinton and Strong[24] recently have repeated these observations obtaining similar values. Sinton and Strong and some of the earlier investigators also were

able to scan the planetary disk latitudinally and longitudinally so as to obtain the diurnal and seasonal variation of temperature as a function of latitude. The up-to-date information can be summarized as follows:

Maximum temperature at equator	~300°K
Mean amplitude of diurnal variation	
Noon to sunset	~60°K
On earth in desert	~30°K
Night side temperature cannot be measured but probably can be estimated at the equator	~200°K
Day side temperature at poles	~220°K
Mean temperature of day side	~260°K
Mean temperature of whole planet	~230°K

The mean temperature of the day side of the planet should be compared with the temperature values obtained by radio measurements at 3-cm wavelength. Mayer, McCullough and Sloanaker[25] found an apparent blackbody disk temperature for Mars of 218°±50°K. Later Giordmaine et al. using Maser techniques,[26] made a more precise measurement of 211°±20°K. According to Mayer,[1] the radiation at radio frequencies probably is emitted from a few centimeters beneath the surface of the planet. As these measurements refer to the day side of the planet, the region just below the surface will be cooler than the surface itself, which will explain the discrepancy between the infrared and radio measurement.

Sinton and Strong[24] also have confirmed the earlier results regarding the difference in temperatures between the dark and bright regions on Mars, the dark region being warmer by ~8°K.

Vertical Distribution of Temperature

The atmosphere of Mars is optically thin in the infrared, and the probable absorbing gases are CO_2 and H_2O, which also are present in Earth's atmosphere. The vertical temperature profile in the atmosphere of Mars, therefore, cannot be determined by observation from the surface of Earth. The measurement is feasible from space vehicles by observing at the center of the strong CO_2 bands, and such experiments may be expected in the future. Only theoretical estimates, therefore, exist regarding the vertical temperature structure of Mars, but the models thus far derived vary considerably from author to author. Urey[10] has summarized and commented upon the earlier results of Hess[27] and Goody[19] and has given his own estimate of the height of the tropopause on Mars, ~30 km.

For the composition of Martian atmosphere, as given in Table 2, the adiabatic temperature gradient will be ~−3.7°K/km. For an equatorial ground temperature of 300°K, Urey therefore obtains a tropopause temperature of 187°K. Goody[19] had previously treated two model atmospheres for Mars, the first in which only CO_2 is the infrared absorber, and the second where 10^{-2} g/cm² of water vapor is responsible for the infrared opacity. The ground temperature was assumed to be 270°K. The tropopause in the two cases was estimated to be at 8.5 and 25 km, respectively. It is not known which of the two models is closer to the real Martian atmosphere. In fact, Jastrow and Rasool[28] recently have shown that, in order to explain the observed mean surface temperature of 230°K by a greenhouse effect, the infrared absorption by the atmosphere must be greater than what can be accounted for even by the inclusion of the specified amounts of both CO_2 and water vapor. If both of these gases are present, the Martian tropopause may lie at an intermediate level between 9 and 25 km.

Arking[4] recently has calculated a model atmosphere for Mars, allowing for convection and using the exact equation of radiative transfer for frequency independent absorption. A total optical thickness of 0.5 was chosen to obtain a surface temperature of 235°K, consistent with the observed mean surface temperature. Assuming an effective blackbody temperature of 217°K, an adiabatic gradient of −3.7°K/km in the convection zone, and an exponential dependence of absorption on altitude with a scale height of 17 km, the temperature profile shown in Fig. 4 (insert) is obtained. The convection zone is found to extend up to 8 km.

Ohring[6] also has investigated recently the vertical temperature profile for a model Martian atmosphere containing 2% CO_2, 98% N_2, and no water vapor. The ground temperature was as-

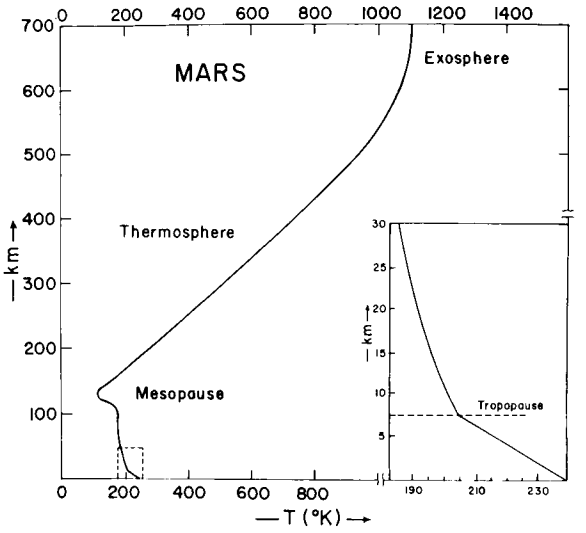


FIGURE 4.—Probable vertical temperature structure of the Martian atmosphere (after Arking, 1962[4] and Chamberlain, 1962[5]).

sumed to be 230°K, and the transfer of radiation in the atmosphere was calculated for frequency-dependent absorption by the CO_2. The tropopause in this case was found to be at 9 km at a temperature of 196°K. In the stratosphere, the temperature keeps on decreasing and reaches a value of as low as 90°K at an altitude of 42 km, where the total pressure is 2.5 *mb*. As pointed out by Ohring himself, at such low stratospheric temperatures the atmospheric CO_2 will freeze out. The author attributes CO_2 cloud layer thus formed to the observed phenomenon of "blue haze" in the Martian atmosphere.

On the basis of arguments developed in a later section, this explanation seems unacceptable, and the stratospheric temperatures in Mars are probably higher than the frost points of CO_2 at the corresponding pressure levels. At 42 km altitude, carbon dioxide will condense at a temperature of 140°K.

The forementioned calculations of the temperature distributions in the Martian atmosphere do not take into account possible heating of the lower atmosphere by direct absorption of solar radiation in the ultraviolet. In the case of the terrestrial atmosphere, the ozone heating produces a temperature maximum at 50 km, but in Mars, with much less oxygen as compared to Earth, ozone would be confined to lower layers of the atmosphere.[28–30] The solar radiation in the region of 2500 Å responsible for ozone dissociation and atmospheric heating will penetrate to a much lower depth in the Martian atmosphere. Because of a higher atmospheric density at this level (compared to the density in Earth's atmosphere at 50 km) and because of the lower intensity of the solar radiation at the distance of Mars, the heating rates probably will not be as important as in the ozonosphere on Earth. However, they may affect the temperature gradients substantially and limit the extent of the convection zone.

Surface Features of Mars

As seen by telescope, the outstanding features of Mars are 1) the dark areas (maria), 2) the general reddish-orange background ("deserts"), and 3) the polar caps.

The nature of the dark maria is not very well understood because, apart from showing fairly regular seasonal changes, they also are subject to erratic variations that make the hypothesis of Earth-type vegetation somewhat doubtful. Moreover, the infrared spectrum does not have characteristic absorption bands of chlorophyll.[12] The presence of organic material in these regions, however, has been reported by Sinton,[31] who observed C–H vibration absorptions in the 3.5 μ region which were absent in the reflection spectrum of the desert areas.

Another factor requiring explanation regarding the dark areas is the higher temperature observed by Sinton and Strong[24] which indicates that on an average they are 8°K warmer than the bright areas. If the "maria" were areas of Earth-type vegetation, one would expect a higher water vapor content than in the desert regions at the same temperature. This in turn would mean that the measurements from Earth in the 8 to 12 μ window would be contaminated by water vapor emission from the atmosphere, and one would expect to see a lower temperature in the darker regions (cf., observations by Tiros in the 8 to 12 μ region of central Africa and the Sahara).[32] The observed higher temperature in the dark areas of Mars is, however, in agreement with

the lower albedo of these regions if interpreted as a strictly surface phenomena.

The probability of the existence of life on Mars is extremely controversial and is beyond the scope of this review on planetary atmospheres, but it will be desirable to see Sinton's observations of C–H bands on Mars repeated and probably substantiated by infared spectroscopic measurements of the terrestrial "dark and bright" areas from an Earth satellite.

The large orange areas of Mars are responsible for the reddish color of the planet seen by the naked eye. According to Kuiper[12] they are composed of felsite rhyolite. Dollfus,[18] from polarization studies, concludes that these areas are predominately limolite, which is pure dehydrated iron oxide. The exact nature is therefore still in debate. De Vaucouleurs[23] has treated this subject in great detail, and no later observational results have been obtained which would modify this discussion.

The polar caps are rather reliably known to be a thin layer of H_2O frost deposited on the surface. Both Kuiper[12] and Dollfus[18] agree on their nature and estimate their thickness as varying between 1 and 5 cm. These caps are observed to expand to lower latitudes during fall and winter and to recede to high latitudes during spring and summer. The evaporation of the polar caps in summer and the almost immediate transport of the water vapor towards the winter pole across the equator are difficult to understand in view of the fact that in Earth's atmosphere the hemisphere to hemisphere mixing is extremely small. De Vaucouleurs[15] has estimated the average speed of this humidity wave across the planet to be $\sim$45 km/day. This pole-to-pole circulation has no parallel on Earth and perhaps can be understood only by the greater length of the Martian year and the increased temperature difference between the cold and warm poles.[15]

Blue Haze Layer

Photographs of Mars taken through filters at wavelengths less than 4500 Å indicate no surface features, and it is therefore believed to be covered with a haze layer known as "blue haze," absorbing at these wavelengths. Sometimes during the opposition, however, it clears over certain regions, and surface features become observable in the blue. These "blue clearings" last usually a few days. Slipher[33] has reported a clearing during the opposition of 1954 lasting for two months. Conflicting arguments have been extended to explain the nature of the blue haze, but an explanation that satisfies all the observed features is yet to come.

Kuiper has suggested that the haze is made of ice crystals of size 0.3 to 0.4 μ located at a height of 6 to 10 km. But, because of the fact that the albedo of Mars decreases very sharply in the blue, the presence of material that absorbs radiation of wavelength less than 4500 Å is required. As water drops or ice crystals are both transparent in the blue, Goody[19] suggests that the particles may be composed of dust of the desired absorbing properties. However, if the particles are nonvolatile, then the explanation of the blue clearing becomes still more difficult. Sharanov,[34] on the other hand, proposes that the blue haze is an optical phenomenon readily explainable by particle and Rayleigh scattering in the lower atmosphere of Mars.

Urey and Brewer[35] have pointed out that CO_2^+, N_2^+, and CO^+ ions absorb in the blue and ultraviolet but fluoresce in the longer wavelengths. The clearings at the opposition then can be explained by the diminution of the solar particle radiation arriving at Mars due to the deviating effect of Earth's magnetic field. Sagan,[36] however, recently has given estimates of the solar particle flux at the distance of Mars and its variation due to Earth's magnetic field and concludes that such mechanism is infeasible.

As mentioned in the previous section, Ohring[6] has suggested that a frozen CO_2 cloud layer at an altitude of $\sim$30 km in the Martian atmosphere may account for the "blue haze" phenomenon. If such were the case, the haze layer would obscure the surface at *all* visual wavelengths,[37] and the absorption would not be confined only in the blue.

Also, if the haze layer were situated at a high level in the Martian atmosphere as suggested by the hypothesis of ions absorption or stratospheric "dry ice" layer, then the apparent diameter of the planet in the blue and in the red should be

considerably different. The very recent measurements of Dollfus[38] indicate a difference of less than 0.1% in the two diameters, thus implying that the level of the haze layer on Mars may not be at an altitude higher than 10 km. As interesting explanation of the "blue haze" recently has been proposed by Warneck and Marmo.[39] It has been shown that a trace amount of NO_2 in the atmosphere of Mars can provide the observed opacity in the blue wavelength region. A more detailed and careful examination of this possibility, however, is desirable.

Clouds

Several types of thin clouds frequently have been observed on Mars. They can be divided into three types:

(1) Blue clouds are visible only in the blue. They are in patches and are seen near the poles and near the terminator. Polarization measurements indicate their particle size to be approximately 0.1 μ[18] and they occur probably at altitudes less than 100 km (see discussion of blue haze). According to Goody,[19] they may be composed of ice crystals formed on the nuclei fed by the haze layer, whereas others believe them to be of the same material as the blue haze.

(2) White clouds are visible in both the blue and yellow light. The polarization studies of these clouds suggest their nature to be the same as ice crystal clouds of a size of approximately 1 μ.[18] Being composed of bigger particles, they probably lie at altitudes less than that of blue clouds.

(3) Yellow clouds, visible only in yellow, are very rare and variable in size. They have been seen drifting several hundred miles across the planet at a velocity of 60 km/hr.[15] According to Goody,[19] they are composed of the same blue absorbing material as the blue haze. The reason why they are not visible in the blue is that they lie above the blue haze, and as the haze is optically thick in the blue and thin in the yellow and red, an overlying cloud of the same material will be seen only in the yellow and red. Sinton and Strong[24] have measured the temperature at the top of these yellow clouds and find a value $\sim$25°K less than the probable ground temperature. Hess,[21] assuming an adiabatic temperature gradient of -3.7°K/km, estimates the height of yellow clouds to be $\sim$6 to 7 km.

Upper Atmosphere and Ionosphere of Mars

Assuming an atmospheric composition of 98% N_2 and 2% CO_2 and a Martian stratospheric temperature of 134°K,[19] Chamberlain[5] recently has computed the mesospheric cooling and thermospheric heating for Mars. With the assumed model atmosphere and from the considerations of CO_2 dissociation into CO and O and the consequent CO cooling at the mesopause, Chamberlain deduces the height of the mesopause as $\sim$130 km at a temperature of 76°K. The considerable CO cooling at the mesopause level acts as "an effective thermostat, keeping the temperature at the exospheric or escape level (1500 km) from exceeding 1100°K."[5] The lifetime of oxygen on Mars according to Eq. (1) turns out to be 10^9 yr,[5] implying its probable retention by the planet. Urey[10] previously had argued that, because of the greater height of the escape level on Mars, thermal conduction to lower layers must be less effective, giving the exospheric temperature of the order of 2000°K. This meant that all gases with atomic or molecular weight <20 certainly would escape from the planet in time $<10^9$ yr. In this exospheric model, however, CO cooling at the mesopause was neglected.

It is difficult to say now rigidly the 1100°K figure applies to Mars in actuality. In the case of Earth's exosphere, it recently has been shown that, though the mean temperature is of the order of 1400°K, the diurnal variations have an amplitude of as much as 500°K.[40] Moreover, short- and long-period variations in the temperature value, highly correlated with the solar activity, also have been observed.[40,41] The daytime exospheric temperature, at solar cycle maximum, may attain a value as high as $\sim$2200°K.[40]

For the considerations of the gravitational escape of gases, the maximum temperature is more relevant. Hence, in the case of Mars, even if one assumes the average exospheric temperature to be 1100°K, the escape of gases will be governed by the day-side maximum temperatures attained during periods of high solar activity. Certainly more oxygen would escape if the

exospheric temperature of Mars occasionally rose to 1500°K (even for a total time of $\frac{1}{10}$ of planetary history) than if it remained steady at 1100°K. The retention of oxygen by Mars is less probable than estimated by Chamberlain.

Figure 4 gives a rough picture of the vertical temperature profile of the Martian atmosphere. The temperatures up to an altitude of 100 km are based on the results obtained by Arking,[4] whereas for above this altitude the upper atmospheric model developed by Chamberlain has been used.

Danilov,[42] Yanow,[43] and Chamberlain[5] have computed the probable electron densities in the ionosphere of Mars, shown in Fig. 5. Danilov has considered an atmosphere composed of N_2^+, NO^+, O^+, and O_2^+. The curve drawn after Chamberlain in Fig. 5 is an upper limit to the possible electron densities but is extremely useful, because it shows the possibility of a secondary maximum to be as low as 130 km. As pointed out by Kellogg and Sagan,[37] Yanow has not considered the importance of electron attachment, and his curve given here shows only O^+ density as derived for 98% N_2 and 2% CO_2 atmosphere, considering a three-body recombination of O with CO or N.

The forementioned ionospheric models are based on the assumption that only the solar ultraviolet and x rays are the ionizing agents in the upper atmosphere. If Mars has a weak magnetic field, the ionization by solar proton flux could be very significant. Also, as the orbit of Mars is near the asteroid belt, a heavier stream of meteoric dust particles is expected to be entering the planet. There is some evidence that Earth's E region is partly maintained by the energy supplied by such meteor streams,[44] and, if the accretion rate of interplanetary dust at

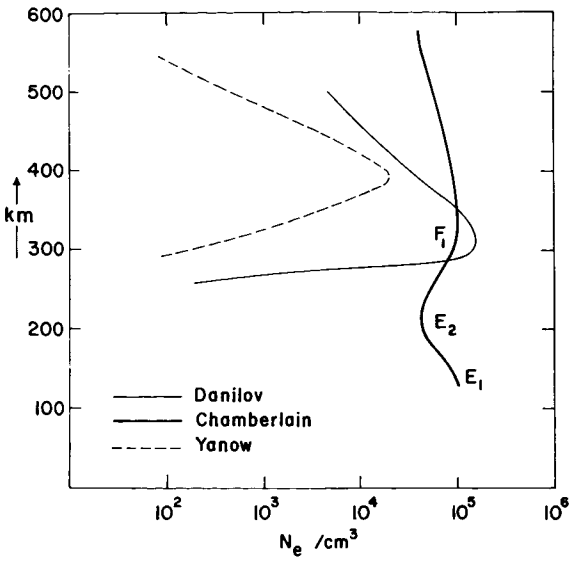


FIGURE 5.—Ionospheric models for Mars (electron density vs. altitude).

Mars is really higher than at Earth, a more important effect on the ionosphere of Mars can be expected. These possibilities are, however, highly hypothetical and have been mentioned only to emphasize the tentative nature of the curves in Fig. 5.

VENUS

Venus is our nearest planetary neighbor and, after the sun and moon, the brightest object in the sky. It therefore has attracted the attention of man since the beginning of civilization. Despite the great interest, very little is known about the atmosphere of this planet, expecially when compared with the information available about Mars. The main reason for this deficiency is that Venus is covered with a layer of white clouds, and the

TABLE 4.—*Astronomical Data for Venus*

	Mass	Rad.	Distance, a.u.	Density	Albedo	T_e, °K	g, cm/sec^2
Earth	1	1	1	5.5	0.39	245	980
Venus	0.81	0.97	0.72	4.8	0.73[a]	235	842

[a] This new value of albedo (cf., 0.76 given by Kuiper (reference 12) and hitherto generally accepted) takes into account the lower albedo of Venus in the infrared recently reported by Sinton (reference 45).

surface remains permanently invisible. Observations have been made, however, in the infrared and radio-frequency regions, and new information regarding the composition and temperature distribution in the Venus atmosphere has been obtained in the past few years. These new results of temperature and pressure at the various levels in the Venus atmosphere have forced a complete revision of ideas regarding the atmospheric structure of this planet. The author will try to summarize the present-day knowledge regarding the Venus atmosphere by giving the physical constants of the planet (table 4) and reviewing the various hypotheses relative to the structure of the Venus atmosphere.

Composition

From the analysis of reflected solar spectrum, the only atmospheric constituent so far detected in Venus is CO_2.[46] Its abundance above the effective "reflecting level" of the 8000-Å photon has been estimated by several workers.[47,48] Until recently, the amount of CO_2 above the cloud level was accepted to be 1000 m-atm,[10] which made up 95% of the total atmosphere above the clouds. Recent reinterpretations of the old spectra of Venus by Spinrad,[2] however, give a CO_2/atmosphere ratio by volume of only 5%, which for a cloud top pressure of 90 *mb* (see discussion on pressure) will correspond to only 40 m-atm of CO_2 above the clouds.

Attempts also have been made to detect water vapor in the atmosphere of Venus. From high altitude balloon measurements, Strong[49] in 1960 obtained a value of 2×10^{-3} g/cm^2 for water vapor above the cloud. The interpretation of these measurements, however, must be questioned in the light of subsequent observations,[50] which indicate a comparable amount of water vapor in Earth's stratosphere. If, however, the clouds of Venus are composed of water drops or of ice crystals,[51,52] a large abundance of water vapor can be expected below the clouds. The amount of water vapor above will be limited by the saturated vapor pressure at the temperature of the cloud tops. Recently Spinrad[2] also has looked for water vapor bands but has failed to detect any.

No other constituent in the Venus atmosphere has been observed spectroscopically, except that recent high resolution measurements of Sinton[45] give a value of 4 cm-atm for CO above the relevant reflecting level of the 2.0 μ photon, but Kuiper,[53] by similar measurements, has not been able to observe any CO bands. Upper limits have been placed on the abundance of other gases that could be expected in the planet's atmosphere.[12] Interpreting the results of Dunham,[16] Urey[10] estimates the amount of oxygen above the clouds to be ~80 m-atm. Again, according to Urey,[10] because of the presence of carbon dioxide, CH_4 and NH_3 cannot be expected to be important constituents of the Venus atmosphere. Also, since all the oxides of nitrogen are unstable with respect to oxygen and nitrogen, they cannot be present in large quantities in any planetary atmosphere.[10] Because it lacks absorption bands in the visible and in the near infrared, nitrogen cannot be detected by Earth-bound observations. Because of its cosmic abundance, however, the presence of N_2 in Venus is most likely, and, in order to account for the pressure, it is generally assumed that the rest of the atmosphere is composed of nitrogen.

Based on Spinrad's estimate of CO_2 mixing ratio, the following composition will be adopted for the Venus atmosphere: N_2 ~95%, CO, ~5%, H_2O ~ indefinite, and O, <80 m-atm above the clouds.

Pressure

There are two rather reliably determined pressure points in the Venus atmosphere:

(1) Dollfus[54] has observed a difference in polarization in the red and green and has interpreted this observation as the result of molecular scattering in an 800-m-thick atmosphere the cloud top. For the slightly lower gravity of Venus, this corresponds to a pressure of 90 *mb* at the cloud top.

(2) De Vaucouleurs and Menzel,[55] by their observations of the occultation of Regulus by Venus, estimate a pressure of 2.6×10^{-3} *mb* at an altitude of ~70 km above the top of the clouds.

Temperature

The temperature measurements are more numerous, but the altitudes in the atmosphere to

which they refer are very uncertain. Sinton and Strong[56] have confirmed the initial 8 to 13 μ radiometric temperature measurements[57] and found a value of 235±10°K. This temperature should originate from a level where the CO_2, if it is the principal radiator in this spectral region, would become optically thick as seen from the top of the atmosphere. The 9.4 and 10.4 μ absorption bands of CO_2 are extremely weak, and, to make CO_2 radiate effectively in this region of the spectrum at relevant pressure and temperature, one must have a depth of at least 400 m-atm. If there are only 40 m-atm of CO_2 above the cloud, the origin of the temperature 235°K is then probably at the cloud top.

From an analysis of the intensity distribution within the CO_2-8000°K-vibration-rotation band levels, Chamberlain and Kuiper[58] estimate a temperature of 285±9°K. The same photographic plates of the Venus spectra obtained in 1932 at Mt. Wilson were recently re-analyzed by Spinrad, and he obtained a rotational temperature varying between 250° and 440°K from plate to plate, i.e., from day to day, which also corresponded to different values of pressure—the low temperatures corresponding to lower pressures, and higher temperature presumably originating from a pressure level up to 6 atm. Spinrad has interpreted these results as probable breaks in the cloud and suggested that the higher temperatures and pressures probably refer to levels far below the visible cloud layer.

De Vaucouleurs and Menzel,[55] while observing the occultation of Regulus and Venus, have deduced fading rate due to differential refraction in the upper atmosphere of Venus and have obtained a scale height $H=kT/mg=6.8$ km and for its first derivative $(1/H)\times(\delta H/\delta z)$, a value of 1%/km. The authors, assuming a mean molar mass of 42.5, estimated a temperature of 297°K and a temperature gradient of 3°K/km at a height of 70 km above the cloud layer. If, however, the assumed atmosphere is 95% N_2, then this assumption of $M=42.5$ is too high. For $M=29$ (for an N_2–CO_2 atmosphere as proposed by Spinrad), one obtains a temperature of ~203°K and a positive temperature gradient of 2°K/km, indicating that the minimum temperature at the mesopause by Venus is probably ~200°K. From the consideration of the change in pressure from 90 *mb* at cloud level to 2.6×10^{-3} *mb* at 70 km above the clouds, one derives a mean scale height of 6.7 km. For a mean molecular weight of 29, this will correspond to a mean temperature of 207°K. This will imply that the temperature in the stratosphere of Venus, for for the most part, remains ~200°K. If the cloud top pressure is ~ 90 *mb* and if 40 m-atm of CO, is the only absorbing gas above it, the problem is very similar to that of the atmosphere of Mars, the cloud level here corresponding to the surface of Mars. The vertical temperature profile in an atmosphere in radiative equilibrium and having a given optical thickness can be estimated approximately from the following simple relation:

$$T_{(z)}{}^4 = T_e{}^4(\tfrac{1}{2}+\tfrac{3}{4}\tau_z)$$

where $T_{(z)}$ is the temperature at the level z, T_e is the effective blackbody temperature of the planet, and τ_z is the optical thickness of the atmosphere above the level z:

$$\tau_z = \int_z^\infty \kappa\rho\, dz$$

Knowing the temperature at the cloud level $T_{(z=0)}$, one can estimate the optical thickness of the atmosphere above the clouds. For Venus, T_e~235°K and $T_{(z=0)}$ at cloud level also (by pure accident) ~235°K; $\tau_{(z=0)}$ then is ~0.7. This implies that the mean transmission in the infrared of the atmosphere above the clouds is $e^{-0.7}$~50%. Now τ_z, at any other altitude above, varies as

$$\tau_z = \tau_{(z=0)} \exp\left[-\int_{z=0}^{z} (dz/H)\right]$$

where H is the scale height ~7 km, and so one can compute the temperature profile for the atmosphere above the cloud level. The temperature attains an asymptotic value of 197°K at ~20 km above the cloud and, because of photodissociative and photoionizing reactions, probably starts increasing again at ~60 km. Figure 6 shows the probable temperature profile of the atmosphere of Venus above the clouds if it is in purely radiative equilibrium.

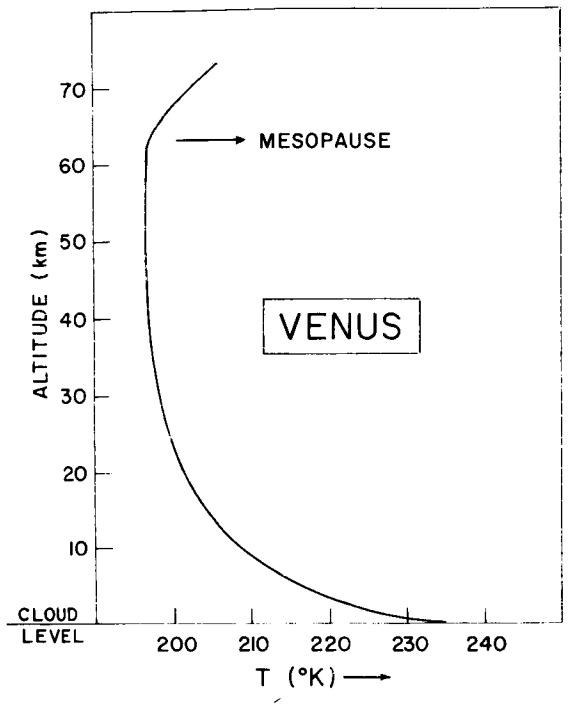


FIGURE 6.—Radiative temperature profile of the Venus atmosphere above the clouds.

It is interesting to compare this result with that obtained by Mintz,[59] who assumed a much greater amount of CO_2 above the clouds and thereby obtained a different temperature profile. The extent of increase in temperature in the ionosphere is not very well known and requires a careful treatment, because cooling processes due to CO, CO_2, and O_2 largely control the temperature at the mesopause.[5]

The structure of the lower atmosphere of Venus is very poorly understood. Urey[10] in 1959 had estimated a ground temperature for Venus of the order of 320°K, but recent measurements of the intensity of the radiation emitted by the planet in the centimeter wavelength show that it corresponds to thermal radiation of temperature of $\sim$600°K! Since radiation in the decimeter region probably passes through the atmosphere and clouds of Venus, it generally is assumed that the measured temperature refers to the surface of Venus. Also, Sagan[52] has shown that the observed temperature spectrum (Fig. 7) cannot be interpreted as synchrotron or cyclotron radiation. The interpretation of these measurements is as yet one of the most important problems in the physics of planetary atmospheres.

Though the observational points are still very few and some of the results are fraught with uncertainties (e.g., variation of measured radio temperatures with phase angle, period of rotation of the planet, tilt of the axis, etc.), several atmospheric models already have been derived in order to explain the high radio-brightness temperatures. Kellogg and Sagan[37] recently have published an excellent review of the proposed model atmospheres for Venus, and therefore they will be described only very briefly, with special emphasis on more recent results.

The temperature measurements shown in Fig. 6 have been made mostly at inferior conjunction of Venus. The high temperature of $\sim$580°K therefore refers to the night side of the planet. As the phase angle decreases, the radio measurements become more and more difficult, i.e., the signal to noise ratio becomes smaller and smaller, making the measurements very uncertain. Mayer[1] has recently summarized the measured temperatures as a function of phase angle. There is some indication of higher than 600°K temperatures for smaller phase angles, but because of the uncertainty of $\sim\pm$100°K in each temperature value, the existence of a phase effect cannot be established. More recently, from the 1961 inferior conjunction measurements, Mayer et al.[60] and Kuzmin and Salomonovich[61] report different magnitudes of the phase effect for different wavelengths, which has

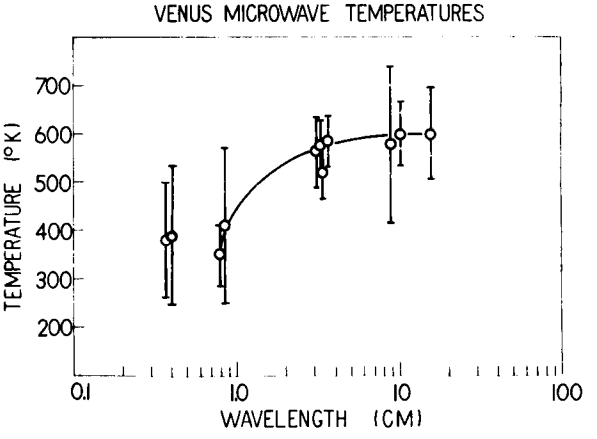


FIGURE 7.—Observed microwave brightness temperature of Venus at different wavelengths (after Mayer, 1961[1]).

made the whole problem still more confusing. The best estimates[60] at present put the dayside temperature at ~750°K, but unless more observations confirm the phase effect, this value should be taken as highly tentative.

Radar Observations

Because of the complete cloud cover, no sharp features are distinguishable to determine the rotation period of Venus by optical measurements. Radar techniques recently have been employed, but the results are very controversial. Victor and Stevens[62] report a period of rotation of 225 days, which is also the period of orbital revolution of Venus, thereby indicating a synchronous rotation of the planet. According to Kotelnikov,[63] however, Venus is rotating much faster, and the period has been estimated to be between 9 and 11 days.

There is also some indication[60] that the minimum radio brightness temperature is not observed at the maximum phase angle, and instead there is a time lag of several days; Kuzmin[61] and Opik[3] therefore argue that the planet may not have a synchronous rotation. If, however, the high radio-brightness temperatures refer to the surface of the planet and the phase effect on the temperature is finally established, a "considerably" slower rotation of the planet is probable.

Despite these uncertainties and lack of comprehensive observational data, the following atmospheric models have been proposed.

Models for Venus Atmosphere

1 *Greenhouse Model*

In this model, it is assumed that the 600°K radio-brightness temperature actually exists at the surface of the planet and is maintained by a very effective "greenhouse effect." The solar radiation (minus albedo) in the visible penetrates up to the surface; the planet, thereby being heated up to a temperature T_e, emits in the infrared, but because of the presence of triatomic molecules like CO_2 and possibly H_2O, which have strong absorption bands in the infrared region, most of the radiation remains trapped in the atmosphere and heats up the surface.

Sagan[52] has estimated the required degree of absorption as 99.5%, which, according to him, will be obtained by an atmosphere composed of 18 km-atm of CO_2 and 10 g/cm^2 of H_2O. From a more elaborate treatment of the problem of radiative transfer in a planetary atmosphere, Jastrow and Rasool[28] recently have pointed out that much higher atmospheric opacity will be required in order to obtain a Venus ground temperature of 600°K. As described previously, T_G (ground temperature) can be obtained from the relation

$$T_g^4 = T_e^4(1+\tfrac{3}{4}\tau_0)$$

Inserting the value of 600°K for T_G and 235°K for T_e, one finds $\tau_0 = 42.5$, which corresponds to a transmission of only $e^{-42.5}$ or $10^{-18.5}$. An absorption of 99.5%, which corresponds to τ_0 of approximately 5, would raise the ground temperature to only 340°K. Thus, if the surface temperature has to be raised to 600°K by a greenhouse effect, the absorption by the atmosphere must be extremely high. Although the model atmosphere proposed by Sagan becomes considerably opaque at high temperatures and pressures, yet even at 600°K and 2 atm pressure the optical thickness in the infrared (τ_0) does not exceed 12.[28] If, therefore, the 600°K surface temperature is attained by a greenhouse effect, then apparently the actual atmosphere of Venus is much different from the model adopted here. Also, recent findings of Spinrad indicate that much less CO_2 is present in the Venus atmosphere than was believed so far.

Water vapor is one of the most effective absorbers of infrared radiation, and large amounts (>100 g/cm^2) at high temperatures certainly can provide infinite optical thickness. Martynov[64] also has proposed a model atmosphere containing large amounts of water vapor, but the absence of water vapor absorption bands in the spectra analyzed by Spinrad[2] (which probably refer to atmosphere below the clouds) makes this model difficult to accept.

If, however, extremely high pressures (~50 atm or more) are prevalent, the pressure broadening of individual absorption lines and induced dipole absorption probably will produce the required opacity even for relatively small amounts of CO_2 and H_2O.[28]

2 Aeolosphere Model

Opik[65] contends that the radiative greenhouse effect cannot account for the surface temperature of 600°K and suggests that the blanketing must be due to dust, wind friction at the surface being the main source of energy for the high temperature. The dust probably is made of calcium and magnesium carbonates, and the atmosphere is composed mainly of CO_2 and N_2. No water vapor is necessarily present in the atmosphere. Because of the blanket of dust, there is no sunlight penetrating to the surface. According to this model, therefore, the radio temperatures, if they refer to the surface, should not show any dependence on the phase angle. The reality of higher temperatures at low phase angles as described previously is still controversial, and more measurements near the superior conjunction of Venus are needed to clarify the situation.

3 Ionosphere Model

Lastly, there is a possibility that the high apparent temperature measured in the centimeter wavelengths region may refer to the ionosphere.[66] In this case, the surface would be at a temperature of ~300°K, whereas the ionosphere would contain a large concentration of free electrons up to a considerable depth, which would lead to a free-free transition and will account for this high temperature. The electron density for the ionospheric thickness of about 300 km (comparable to Earth) for such a model is about $10^9/cm^3$. This is about 1000 times greater than the maximum density in Earth's ionosphere. It is very difficult to envisage such high electron concentration unless an ionization mechanism is found which is many orders of magnitude more effective than the solar ultraviolet and x rays. The solar proton flux has been suggested[66] as the other source, but, from its value as known at the distance of Earth and assuming that Venus has a very weak magnitude field, this would provide an ionization mechanism at the maximum of only ~30 times more effective than the ultraviolet radiation.[37]

At the moment, no known mechanism can provide such high electron densities as $10^9/cm^3$ in the Venus ionosphere, and, moreover, there is another fact observed which argues against the ionospheric model. If the ionosphere is optically thick at $\lambda = 3$ cm, it certainly will be opaque at 12.5 cm, which is the wavelength used for the radar measurements. Now, as the ionosphere cannot be opaque and reflect at the same time, the electron density should be still higher ($\sim 10^{12}/cm^3$) in order for the ionosphere to reflect at 12.5-cm wavelength. It also has been suggested that there is an ionospheric hole at the midnight point (or antisolar point) of Venus,[37] and, as the radar measurements have been made only at the inferior conjunction, this may account for the radar reflections. These suggestions are, however, highly speculative, and an ionosphere of such high electron density in the first place and its failure to account for the radar reflectivities make the ionospheric model for the explanation of 600°K temperature very improbable.

Apart from the high temperature observed at wavelengths >3 cm, there is another aspect of the measured temperature values (Fig. 7) which requires explanation: the apparent decrease in the brightness temperatures at wavelengths < 1 cm.

The forementioned atmospheric models give a tentative explanation of the low temperature obtained at 8 mm. In the case of the greenhouse and aeolospheric models, the radiation corresponding to 350°K as observed at 8 mm is emitted from the middle of the troposphere, whereas the ionospheric model, in which the ionosphere is emitting at 600°K, the lower temperature refers to the surface of the planet.

Barrett[67] has shown that self-absorption of an atmosphere composed of CO_2 and H_2O at a total pressure of ~20 atm and surface temperature of 600°K will account for the observed temperature spectrum. Rasool,[22] on the other hand, attributes this sudden decrease of temperature at 8 mm to 50% attenuation of microwave radiation at this wavelength by a 2-km-thick cloud layer, assuming that the clouds of Venus were of the terrestrial type made of water with a drop size of less than 50 μ. These models were conceived before the 4-mm temperature measurements had been made, and, except for the ionosphere model, these explanations required a still lower temperature at 4-mm wavelength.

According to Fig. 7, however, the temperatures measured at 4 mm are of the same order of magnitude as at 8 mm. If the level of origin of the radiation at the two wavelengths is the same, then the forementioned interpretations become questionable. The accuracy of temperature measurements at 4 and 8 mm is, however, so low that, before drawing any definite conclusions, more observational points have to be obtained which would give a definite shape to the temperature spectrum. More measurements in the wavelength interval 0.1 to 3.0 cm will be extremely rewarding.

The confused state of knowledge regarding the atmosphere of Venus as outlined previously probably will not last very long. The data relayed back by Mariner II[68] may contain useful clues regarding the physical conditions prevalent at the surface and in the atmosphere of Earth's "sister" planet, which appears to be so different.

JUPITER

Because of their considerable distances from Earth and the probable presence of a thick atmosphere, relatively little is known about the structure of the atmospheres of the major planets. Jupiter is the biggest of all the planets, with a mass $\sim$300 times greater than Earth. As its volume exceeds Earth's by a factor of 1000, the mean density is comparatively small (1.33 g/cm^3, Earth = 5.5 g/cm^3). Being 5 times farther removed from the sun than Earth but having a comparable albedo of 0.47, purely physical considerations indicate that the temperature of the planet should be very low. With a large surface gravity (2.6 times that of Earth) and a lower temperature, one would expect the chemical composition of the atmosphere to be still primitive, containing large quantities of hydrogen and helium. An estimate on the basis of Eq. (1) gives the time of escape of hydrogen from Jupiter of the order of 10^{116} yr. This excludes the possibility of the gravitational escape of any atmospheric gas from the planet and indicates a predominantly hydrogen atmosphere. Recent theoretical and experimental evidence is, however, against this composition of the Jovian atmosphere, and the author will therefore consider the most recent observational results in more detail and attempt to understand the structure of the atmosphere of Jupiter.

Composition

The only gases spectroscopically detected are CH_4 and NH_3. Their abundances, according to Kuiper,[12] are $CH_4 = 150$ m-atm and $NH_3 = 7$ m-atm above the cloud surface. As both hydrogen and helium are undetectable by spectroscopic measurements in the visible, no direct evidence of their presence in Jupiter's atmosphere was available until very recently, when Kiess, Corliss, and Kiess[69] detected the quadrupole rotation-vibration lines of molecular hydrogen in the Jovian spectra. Zabriskie[70] analyzed these spectra to obtain a total hydrogen amount of 5.5 km-atm of molecular hydrogen above the cloud level. This, however, is in complete disagreement with the hitherto accepted atmospheric composition of Jupiter which, according to Urey,[10] is as follows:

Gas	km-atm
H_2	270
He	56
CH_4	0.15
NH_3	0.007

The hydrogen and helium abundances had been derived on the assumption that they are present in solar proportions. In that case, neon and nitrogen also will be present in small quantities. The mean molecular weight of this atmosphere would be 3.25 and the pressure at the cloud layer $\sim$8 bars.

Very recently Opik[3] has questioned this composition because of the following reasons: Baum and Code[71] observed the occultation of σ Arietis by Jupiter. From the rate of fading of the occulated star, they calculated the density scale height $H = kt/mg \sim 8.3$ km for the atmosphere of a planet, and, from this observed value of scale height and the polychromatic radiative equilibrium temperature value of 112°K, Opik[65] estimates the mean molecular weight of the atmosphere of Jupiter above the clouds to be 4.3 ± 0.5, which is considerably different from the value obtained for the previously quoted abundances. Opik therefore accepts the measured hydrogen abundance of 5.5 km-atm[70] and proposes the fol-

lowing composition for Jupiter:

Gas	Percentage
He	97.2000
H_2	2.3000
Ne	0.3900
CH_4	0.0630
Ar	0.0642
NH_3	0.0029

which will give a mean molecular weight of 4.02, in good agreement with the occultation value. Opik also argues against the presence of free nitrogen in the atmosphere which will combine immediately with hydrogen to form NH_3. CO also will be absent, having been reduced to CH_4 and H_2O. Water would remain beneath the clouds and therefore would be undetectable from Earth. If this composition is accepted the fractionation of hydrogen in the earlier history of the planet has yet to be explained.

Temperature

The effective blackbody temperature of Jupiter for a visual albedo of 0.47 is 105°K. Measurements have been made in the 8 to 12 μ window by Menzel et al.[72] which indicate a value of ~130°K. This measurement may not refer to the cloud top level of Jupiter, as both NH_3 and CH_4 emit in the far infrared, and the measured radiation probably may be weighted heavily with atmospheric emission.[73] The measurements at radio frequencies, however, give a very different result. Table 5 summarizes the up-to-date observational results at different frequencies.[1]

It is believed that the radiation observed at ~3 cm wavelength is thermal, but the level to which it refers is not known because it is probably heavily weighted by the pressures broadened 1.28-cm line of NH_3.[73] The highly intense radiatation observed at wavelength >10 cm is probably nonthermal originating from the Van Allen-type belts around Jupiter.[74] A brief discussion of these radiation belts will be given in a later section.

TABLE 5.—*Observed Radio Brightness Temperatures of Jupiter*

Wavelength, cm	Temperature, °K
3.03	171
3.17	173
3.36	189
3.75	210
10.20	640
21.00	2,500
22.00	3,000
31.00	5,500
68.00	70,000

Clouds

Telescopic view of Jupiter shows banded cloud structure. The period of rotation being very smal ($9^h\ 55^m\ 28^s$), the banded cloud layers can be explained easily. The nature of these clouds and the color effects observed in these cloud belts are, however, still controversial and recently have been discussed in detail by Newburn.[73] The clouds probably are made of NH_3, which is quite comprehensible in view of the prevalent low temperatures and high pressure. The observed motion of the clouds, however, poses an interesting problem of meteorology. The equatorial clouds have a period of rotation ~5 min less than those at higher latitudes. Kuiper[12] has given a tentative explanation of this difference in rotation period by assuming a gaseous ring around the planet which would accelerate the equatorial clouds. The problem is far from being resolved and probably is related to the possible differences in the tropopause heights at the equator and at the poles.[75]

Red Spot

Another problem of interest in the atmosphere of Jupiter is the great red spot, which is ~40,000 km in length, ~13,000 km in width, and was first observed in 1831. This spot reached its highest intensity in 1880, when its color became pink. Since then, both visibility and color have waxed and waned. Moreover, it is not attached rigidly to the surface, as it has been observed to oscillate at random. The most common explanation is that of a large meteorite floating on liquid or in a heavy atmosphere, but the changes in color and intensity cannot be accounted for adequately by this hypothesis. Recently, Hide[76] has advanced an explanation in which the surface of the planet in that region would be a plateau only a few kilometers high. The hydrodynamic theory of circulation suggests that, because Jupiter rotates so rapidly, the effect of a shallow topographical feature on the general circulation of the atmos-

phere will be attenuated very slowly with height. Thus, the feature will make its presence manifest at the level of the cloud. This explanation seems quite plausible, except that it is difficult to envisage the presence of only one topographical feature of this size and the absence of other shallow topographical features creating other spots that have not been observed. Moreover why should it have appeared suddenly in 1831?

Radiation Belts

Considerable interest has developed in the origin of centimetric radiation from Jupiter (table 5) which corresponds to extremely high temperatures. Earlier investigations by Field[74] led to the possibility that the observed decimetric radiation was cyclotron and was originating from electrons trapped in a 1200-gauss polar magnetic field. Recently Roberts and Huguenin[77] have observed variation of percentage polarization with solar activity, thus eliminating cyclotron radiation as the mechanism responsible for the Jovian decimetric emissions. These measurements are more consistent with synchrotron radiation originating from highly relativistic electrons trapped in the Van Allen-type belts of Jupiter. These radiation belts presumably are highly populated with energetic particles to a distance of 3 Jupiter radii,[77] and the surface magnetic field is of the order of 50 gauss.

Another source of radiation has been observed in the decameter wavelength region and has been found to be very erratic. The spectrum of this radiation extends from 10 to 21 m, with a maximum around 15.8 m.[78] It has been suggested that this radiation may be coming from the atmosphere or from the surface but is modified by the ionosphere in the presence of magnetic field of about 7 gauss.[79] Warwick,[80] however, explains this radiation in terms of precipitation of fast electrons out of Jupiter's radiation belts and down to the surface of the planet along dipole lines of force. This explanation also will be consistent with the synchrotron model of the decimeter radiation.

REFERENCES

1. Mayer, C. H., "Radio emission of the moon and planets," *Solar System III (Planets and Satellites)*, edited by G. P. Kuiper and B. N. Middlehurst (University of Chicago Press, Chicago, 1961), Chap. 12, p. 442.
2. Spinard, H., "Spectroscopic temperature and pressure measurements in the Venus atmosphere," Publ. Astron. Soc. Pacific **74,** 187 (1962).
3. Opik, E. J., "Jupiter: chemical composition, structure and origin of a giant planet," Icarus **1,** 200 (1962).
4. Arking, A., "Non-grey planetary atmospheres," Onzième Colloque International d'Astrophysique, "La Physique des Planètes," Université de Liège (July 1962).
5. Chamberlain, J. W., "Upper atmosphere of the planets," Astrophys. J. **136,** 582 (1962).
6. Ohring, G., "A theoretical estimate of the average vertical distribution of the temperature in the Martian atmosphere," Contract NASw-286, Geophys. Corp. America TR 2, Bedford, Mass. (June 1962).
7. Jastrow, R., and Kyle, L., "The upper atmosphere," *Handbook of Astronomical Engineering*, edited by H. H. Koelle (McGraw-Hill Book Co., Inc., New York, 1961), Chap. 2, p. 2–2.
8. Dzhasybekova, E. K., Kazachevskii, V. M., and Kharitonov, A. V., "Determination of the earth's albedo," Astron. Zh. **37,** no. 1, 131 (1960); also Soviet Astron—AJ **4,** 125 (1960).
9. Spitzer, L., Jr., "The terrestrial atmosphere above 300 km," *Atmospheres of the Earth and Planets*, edited by G. P. Kuiper (University of Chicago Press, Chicago, 1952), Chap. 7, p. 211.
10. Urey, H. C., "The atmospheres of the planets," *Hanbuch der Physik* (Springer Verlag, Berlin, 1959), Vol. 52, p. 363.
11. Bourdeau, R. E., Whipple, E. C., Jr., Donley, J. L., and Bauer, J., "Experimental evidence for the presence of helium ions based on Explorer VIII satellite data," J. Geophys. Research **67,** 467 (1962).
12. Kuiper, G. P., "Planetary atmospheres and their origin," *Atmospheres of the Earth and Planets*, edited by G. P. Kuiper (University of Chicago Press, Chicago, 1952), Chap. 12, p. 306.
13. Grandjean, J. and Goody, R. M., "The concentration of CO_2 in the atmosphere of Mars," Astrophys. J. **121,** 548 (1955).
14. Kiess, C. C., Corliss, C. H., Kiess, C. H., and Corliss, E. L. R., "High dispersion spectra of Mars," Astrophys. J. **126,** 579 (1957).
15. de Vaucouleurs, G., "Physical environments of Mars," *The Physics and Medicine of the Atmosphere and Space*, edited by O. O. Benson and H. Stronghold (John Wiley and Sons, New York, 1960), Chap. 39, p. 586.
16. Dunham, T., Jr., "Spectroscopic observations of planets at Mount Wilson," *The Atmospheres of the Earth and Planets*, edited by G. P. Kuiper (University of Chicago Press, Chicago, 1952), p. 288.
17. Sinton, W. M., "An upper limit to the concentration of NO_2 and N_2O_4 in the Martian atmosphere," Publ. Astron. Soc. Pacific **73,** 125 (1961).
18. Dollfus, A., "Etudes des planèts par la polarisation de leur lumière," Ann. Astrophys. Suppl. 4 (1957).
19. Goody, R. M., "The atmosphere of Mars," Weather **12,** 3 (1957).

20. ADEL, A., Astrophys. J. **93,** 397 (1941).
21. HESS, S. L., "Mars as an astronautical objective," *Advances in Space Science and Technology*, edited by F. I. Ordway (Academic Press, New York, 1961), Vol. 3, p. 151.
22. RASOOL, S. I., "Infrared opacity of the Venus atmosphere," Am. Geophys. Union Symposium, Univ. Calif., Los Angeles (December 1961).
23. DE VAUCOULEURS, G., *Physics of the Planet Mars* (Faber and Faber, London, 1954).
24. SINTON, W. M. and STRONG, J., "Radiometric observations of Mars," Astrophys. J. **131,** 459 (1960).
25. MAYER, C. H., MCCULLOUGH, T. P., and SLOANAKER, R. M., "Observations of Mars and Jupiter at a wavelength of 3.15 cm," Astrophys J. **127,** 11 (1958).
26. GIORDMAINE, J. A., ALSOP, L. E., TOWNES, C. H., and MAYER, C. H., "Observations of Jupiter and Mars at 3 cm wavelength," Astrophys. J. **64,** 332 (1959).
27. HESS, S. L., "Blue haze and vertical structure of Martian atmosphere," Astrophys. J. **127,** 743 (1958).
28. JASTROW, R. and RASOOL, S. I., "Radiative transfer in the atmospheres of Mars and Venus," *Space Research III* (North-Holland Publishing Co., Amsterdam, in press).
29. MARMO, F. F. and WARNECK, P., "Photochemical process in the atmosphere of Mars," Rept. of Contract NASw-124, Geophys Corp. America, Bedford, Mass. (1961).
30. WILDT, R., "Ozon und sauerstoff in den planeten-atmosphären," Veroffendichungen der Universitäts, Sternwarte in Gottingen, Rept. 38 (1943).
31. SINTON, W. M., "Further evidence of vegetation on Mars," Science **130,** 1234 (1959).
32. NORDBERG, W., BANDEEN, W. R., CONRATH, B. J., KUNDE, B., and PERSANO, I., "Preliminary results of radiation measurements from the Tiros III meteorological satellite," J. Atmospheric Sci. **19,** 1 (1962).
33. SLIPHER, E. C., Telescope **14,** 361 (1954).
34. SHARANOV, V. V., "On the role of true absorption in the Martian atmosphere," Astron. J. USSR **34,** 547 (1957).
35. UREY, H. C. and BREWER, A. W., "Fluorescence in planetary atmospheres," Proc. Roy. Soc. (London) **A241,** 37 (1957).
36. SAGAN, C., "Is the Martian haze produced by solar protons?," Icarus **1,** 70 (1962).
37. KELLOGG, W. W. and SAGAN, C., "The atmospheres of Mars and Venus," Panel on Planetary Atmospheres of Space Science Board, Natl. Acad. Sci.-Natl. Research Council, Washington, D. C., Publ. 944 (1961).
38. DOLLFUS, A., private communication (1962).
39. WARNECK, P. and MARMO, F. F., "NO_2 in the Martian atmosphere," to be published.
40. HARRIS, I. and PRIESTER, W., "Time dependent structure of upper atmosphere," J. Atmospheric Sci. **19,** 286 (1962).
41. JACCHIA, L. G., "A variable atmosphere-density model from satellite accelerations," Smithsonian Astrophys. Obs., Special Rept. 39, Cambridge, Mass. (March 1961).
42. DANILOV, A. D., "Model of Venus and Mars ionospheres," Geomagnetism i Aeronomiya I, 314–319 (1961).
43. YANOW, G., "A study of the Martian upper atmosphere and ionosphere," J. Astronaut. Sci. VIII, 103 (1961).
44. LOVELL, A. C. B., *Meteor Astronomy* (Clarendon Press, Oxford, 1954).
45. SINTON, W. E., "Recent infrared observations of Venus," Onzième Colloque International d'Astrophysique, "La Physique des Planètes," Université de Liège (July 1962).
46. ADAMS, W. S. and DUNHAM, T., JR., "Absorption bands in the infrared spectrum of Venus," Publ. Astron. Soc. Pacific **44,** 243 (1932).
47. HERZBERG, G., "Laboratory absorption spectra obtained with long paths," *The Atmospheres of the Earth and Planets*, edited by G. P. Kuiper (University of Chicago Press, Chicago, 1952), Chap. 13, p. 406.
48. KUIPER, G. P., Publ. Astron. Soc. Pacific **62,** 133 (1950).
49. STRONG, J., private communication (1960).
50. MURCRAY, D. G., MURCRAY, F. G., WILLIAMS, W. J., and LESLIE, F. E., "Water vapor distribution above 90,000 feet," J. Geophys. Research **65,** 3641 (1960).
51. LYOT, B., "Polarisation de la planète Vénus," Ann. Observ. Paris-Meudon **8,** 76 (1929).
52. SAGAN, C., "The radiation balance of Venus," Jet Propulsion Lab. TR 32–34, NASA Contract NASw-6, Calif. Inst. Tech., Pasadena, Calif. (1960).
53. KUIPER, G. P., "Infrared spectra of planets and cool stars," Onzième Colloque International d'Astrophysique, "La Physique des Planètes," Université de Liège (July 1962).
54. DOLLFUS, A., "Polarization studies of planets," *Solar System III* (*Planets and Satellites*), edited by G. P. Kuiper and B. N. Middlehurst (University of Chicago Press, Chicago, 1961), Chap. 9, p. 343.
55. DE VAUCOULEURS, G. and MENZEL, D. H., "Final report on the occultation of regulus by Venus," Air Force Cambridge Research Lab. Sci. Rept. 1, Air Research and Dev. Command Contract AF19,604–7461, Bedford, Mass. (1961).
56. SINTON, W. M. and STRONG, J., "Radiometric measurements of Venus," Astrophys. J. **131,** 470 (1960).
57. PETTIT, E. and NICHOLSON, S. B., "Radiation measures on the planet Venus," Publ. Astron. Soc. Pacific **36,** 269 (1924).
58. CHAMBERLAIN, J. W. and KUIPER, G. P., "Rotational temperature phase variation of the carbon dioxide bands of Venus," Astrophys. J. **124,** 399 (1956).
59. MINTZ, Y., "Temperature and circulation of the Venus atmosphere," Planetary Space Sci. **5,** 141 1961.
60. MAYER, C. H., MCCULLOUGH, T. P., and SLOANAKER, R. M., "3.15 cm observations of Venus in 1961," Onzième Colloque International d'Astrophysique, "La Physique des Planètes," Université de Liège (July 1962).

61. KUZMIN, A. D. and SALOMONOVICH, A. E., "Some conclusion concerning the physical conditions on Venus from radio astronomical observations carried out at the Lebedev Physical Institute," Onzième Colloque International d'Astrophysique, "La Physique des Planètes," Université de Liège (July 1962).
62. VICTOR, W. K. and STEVENS, R., "Exploration of Venus by radar," Science **134,** 46 (1961).
63. KOTELNIKOV, V. A., "Radar contact with Venus," Brit. Inst. Radio Engrs. J. **22,** 293 (1961).
64. MARTYNOV, D. Y., "The physical conditions in the atmosphere of Venus," Onzième Colloque International d'Astrophysique, "La Physique des Planètes," Université de Liège (July 1962).
65. OPIK, E. J., "The aelosphere and atmosphere of Venus," J. Geophys. Research **66,** 2807 (1961).
66. JONES, D. E., "The microwave temperature of Venus," Planetary Space Sci. **5,** 166 (1961).
67. BARRETT, A. H., "Microwave absorption and emission in the atmosphere of Venus," Astrophys. J. **133,** 281 (1961).
68. BARRETT, A. H., COPELAND, J., JONES, D. E., and LILLEY, A. E., "Objectives of the mariner Venus microwave radiometer experiment," Jet Propulsion Lab. TR 32–156, Calif. Inst. Tech., Pasadena, Calif. (1961).
69. KIESS, C. C., CORLISS, C. H., and KIESS, H. K., "High dispersion spectra of Jupiter," Astrophys. J. **132,** 221 (1960).
70. ZABRISKIE, F. R., "The hydrogen content of Jupiter's atmosphere," Astron. J. **67,** 169 (1962).
71. BAUM, W. A. and CODE, A. D., "A photometric observation of the occultation of σ Arietis by Jupiter," Astron. J. **58,** 108 (1953).
72. MENZEL, D. H., COBLENTZ, W. W., and LAMPLAND, C. O., "Planetary temperatures derived from water cell transmission," Astrophys. J. **63,** 177 (1926).
73. NEWBURN, R. L., "The exploration of Mercury, the asteroids, the major planets and their satellite systems and Pluto," *Advances in Space Science and Technology,* edited by F. I. Ordway (Academic Press, New York, 1961), Vol. 3, p. 195.
74. FIELD, G. B., "The source of radiation from Jupiter at decimeter wavelengths 2. Cyclotron radiation by trapped electrons," J. Geophys. Research **65,** 1661 (1960).
75. HESS, S. L., "Variations in the atmospheric absorption over the disc of Jupiter and Saturn," Astrophys J. **118,** 151 (1953).
76. HIDE, R., "Origin of Jupiter's great red spot," Nature **190,** 895 (1961).
77. ROBERTS, M. S. and HUGUENIN, G. R., "The radiation belts of Jupiter," Onzième Colloque International d'Astrophysique, "La Physique des Planètes," Université de Liège (July 1962).
78. CARR, T. D., "Radio frequency emission from the planet Jupiter," Astron. J. **64,** 39 (1959).
79. SMITH, A. G. and CARR, T. D., "Radio-frequency observations of the planets in 1957–1958," Astrophys. J. **130,** 641 (1959).
80. WARWICK, J. W., "The relation between Jupiter's decametric emission and radiation belts," Onzième Colloque International d'Astrophysique, "La Physique des Planètes," Université de Liège (July 1962).
81. DE VAUCOULEURS, G., private commuinication (1962).

GENERAL BIBLIOGRAPHY

BURKE, B. F., "Radio observations of Jupiter I," *Planets and Satellites* (*Solar System III*), edited by G. P. Kuiper and B. M. Middlehurst (University of Chicago Press, Chicago, 1961), Chap. 13, p. 473.

GALLET, R., "Radio observations of Jupiter II," *Planets and Satellites* (*Solar System III*), edited by G. P. Kuiper and B. M. Middlehurst (University of Chicago Press, Chicago, 1961), Chap. 14.

HESS, S. L., "Mars as an astronautical objective," *Advances in Space Science and Technology,* edited by F. I. Ordway (Academic Press, New York, 1961), Vol. 3, p. 151.

JASTROW, R., "Results of experiments in space," J. Aerospace Sci. **29,** 377 (1962).

KELLOGG, W. W. and SAGAN, C., "The atmospheres of Mars and Venus," Panel on Planetary Atmospheres of Space Science Board, Natl. Academy Sci.-Natl. Research Council, Washington, D. C., Publ. 944 (1961).

KUIPER, G. P., "Planetary atmospheres and their origin," *Atmospheres of the Earth and Planets,* edited by G. P. Kuiper (University of Chicago Press, Chicago, 1952), Chap. 12.

MITRA, S. K., "The upper atmosphere," Asiatic Soc. Calcutta (1952).

MOORE, P., *The Planet Venus* (Macmillan Co., New York, 1961), 3rd ed.

NEWBURN, R. L., "The exploration of Mercury, the asteroids, the major planets and their satellite systems and Pluto," *Advances in Space Science and Technology,* edited by F. I. Ordway (Academic Press, New York, 1961).

NICOLET, M., "Effets de l'ultra-violet lointain solarire sur l'atmosphere de la terre et des autres plantes," Mem. Soc. Roy. Sci. Liege Ser. 5 **IV,** 319 (1961).

RATCLIFFE, J. A. (ed.), *Physics of the Upper Atmosphere* (Academic Press, New York and London, 1960).

SAGAN, C., "The planet Venus," Science **133,** 849 (March 24, 1961).

UREY, H. C., "The atmospheres of the planets," *Hanbuch der Physik* (Springer Verlag, Berlin, 1959), Vol. 52, p. 363.

VAN DE HULST, H. C., "Scattering in the atmospheres of the earth and planets," *The Atmospheres of the Earth and Planets,* edited by G. P. Kuiper (University of Chicago Press, Chicago, 1952).

DE VAUCOULEURS, G., *Physics of the Planet Mars* (Faber and Faber, London, 1954).

THE ATMOSPHERES OF MARS, VENUS, AND JUPITER

S. I. RASOOL AND R. JASTROW

Goddard Space Flight Center

The general properties of the atmospheres of Mars, Venus, and Jupiter are reviewed. The evolutionary processes of planetary atmospheres are outlined as an introduction to a detailed discussion of the structure of the atmospheres of other plantes.

The most recent observational results indicate that the surface pressure on Mars may be about 20 mb, and the abundance of CO_2 in the atmosphere of Mars about 25% by volume. For Venus there are indications that the surface pressure is about 100 atm., with the atmosphere mainly composed of N_2. In the case of Jupiter, recent theoretical investigations of the atmosphere below the clouds suggest that there may be regions in the atmosphere of very high density and pressure but at relatively low temperature, creating conditions possibly favorable to the occurrence of complex organic reactions.

THE STRUCTURE OF A PLANETARY ATMOSPHERE

Composition, temperature and pressure are the three parameters which determine the structure of a planetary atmosphere.

Composition

In order to discuss the composition of a planetary atmosphere, we shall go quite far back into the history of our subject and deal first with the processes by which we believe the planetary atmospheres have evolved. By this approach we can hope to gain an understanding of the conditions which limit the structure and composition of the atmospheres of different planets.

If we assume that the planets condensed from a contracting solar nebula, then the initial composition of the planetary atmospheres would be, in general, governed by the relative abundances of elements in the sun. These are now fairly well known and the abundances compiled by Aller[1] are plotted in figure 1.

The peaks in the curve reflect the stability of the respective nuclear species, as is expected for equilibrium concentrations produced by nucleosynthesis. In addition to H and He, the C, N, O group is strong; the Mg, Al, Si group is also strong, with Si the strongest, and Fe is very high. The heavier-than-iron elements fall off rapidly by 3 or more orders of magnitude, with fractional concentrations running about 10^{-9} or 10^{-10}.

These abundances give an indication of the composition to be expected in the contracting cloud which formed the primitive sun. Apart from hydrogen and helium, which should be the most abundant constituents, H_2O, perhaps NH_3

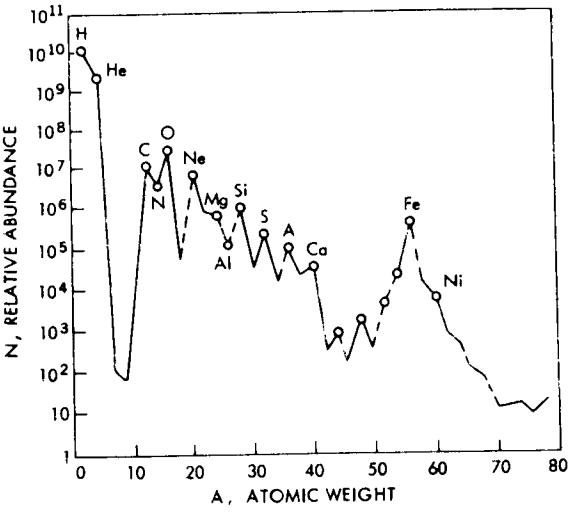


FIGURE 1.—Abundance of the elements in the sun (reference 1): A is the atomic weight and N is the relative abundance compared to $Si = 10^6$.

and CH_4, oxides of C, N, Mg, Si, and Fe, all loosely joined in complex and irregular chains, are expected to form macromolecules and dust particles. It is presumably aggregates such as these which provided the original material of the planets. On this basis, it is understandable that the composition of the earth should be dominated by silcon oxides plus oxides and carbonates of Mg and Ca.

From these considerations it also seems very probable that the initial composition of the planetary atmospheres will be dominated by H_2, He, CH_4, NH_3, H_2O, CO_2, N_2, etc.

Given this primitive composition, the atmospheres will then have evolved through four primary processes:

1. The gravitational escape of gases into the interplanetary space.
2. The exhalation of gases from the interior of the planets.
3. Chemical reactions of gases in the atmosphere with the solid material of the planet.
4. The photodissociation and ionization of gases by the solar radiation.

Also contributing secondarily are the radioactive decay products and the capture of gaseous constituents from the interplanetary medium.

At any level in an atmosphere a number of atoms and molecules will be moving upward with speeds in excess of that required for escape from the gravitational attraction of this plannet. The critical escape velocity, v_c, from a planet is

$$v_c = (2MG/R)^{1/2},$$

where M is the mass of planet, R its radius, and G the universal gravitational constant. For the earth,

$$v_c = 11.3 \text{ km/sec};$$

and for other planets,

$$v_c = 11.3(M/R)^{1/2} \text{ km/sec}$$

where M and R are in units of the earth's mass and radius. For the moon, Mars, Venus, and Jupiter the values of v_c are 2.3, 5.0, 10.4, and 61.0 km/sec, respectively.

The escape of gases can only take place effectively at high altitudes where the density and the probability of collisions are small. These requirements define the region of the atmosphere known as the *exosphere*, a region in which the mean free path for collisions is very large—for a particle moving outward the probability of collision is only one-half. Atmospheric particles in the exosphere therefore execute ballistic trajectories in the gravitational field of the planet.

Knowing the temperature and density at the base of the exosphere, the number of particles can be calculated which will be escaping from the planet per unit time. The classical formula (e.g., reference 2) for the time of escape t_e in which the density of an atmospheric constituent of molecular or atomic weight 'm' will fall to e^{-1} of its original value is given by

$$t_e = [B(6\pi)^{1/2}\bar{v}/3g](e^y/y), \qquad (1)$$

where $B = n_0T_0/n_cT_c$, $y = mv_c^2/2kT_c$, n_0 and T_0 are the particle number density and temperature respectively at the ground level, n_c the hypothetical number density at ground if the whole atmosphere was isothermal at the escape level temperature T_c, $\bar{v}$ the root mean square velocity of the escaping particles, and v_c the velocity of escape.

In the case of the earth, the average temperature of the exosphere has recently been determined to be about 1650°K. For this value of the exospheric temperature, the time of escape of hydrogen from earth is approximately 1000 years. As the age of the planets is about 5 billion years, the absence of hydrogen in the atmosphere of the earth is comprehensible. Gases like CH_4 and NH_3, which are readily dissociated by solar ultraviolet radiation, would also lose their hydrogen and their relative scarcity in the atmosphere is therefore understandable.

However water is still present on the Earth. Due to the low temperature of the tropopause, almost all the terrestrial water vapor remains confined to the troposphere. The extreme ultraviolet which can dissociate water vapor does not reach the troposphere because it is being attenuated by ionizing and dissociative processes at the top of the atmosphere. An extremely minute amount of water vapor, corresponding to the saturated vapor pressure of water at 200°K, will traverse the tropopause and will be susceptible to dissociation in the upper atmosphere. This special circumstance preserves the water on the earth.

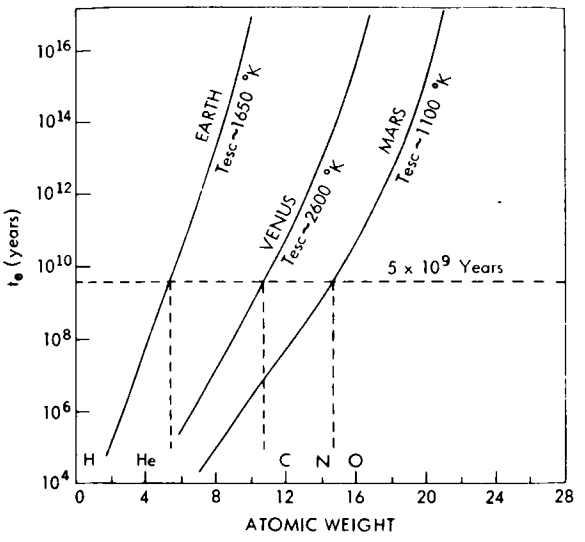


FIGURE 2.—Effective time of escape of gases as a function of atomic weight for Earth, Venus, and Mars.

For helium the time of escape from the earth calculated from equation 1 will be of the order of 10^7 years. The equilibrium amount of helium in the atmosphere at any time will, however, depend both on the rate of outgassing of helium from the crust and the rate of its escape from the exosphere. Such calculations indicate that the amount of helium in the earth's atmosphere should be higher than observed. This discrepancy is not for the present very well understood.[3]

From the above discussion it is clear that atmospheric constituents of molecular or atomic weight greater than 6 have probably been retained by the earth up to the present time, but, as almost all hydrogen must have escaped, the present atmosphere of the earth should be in a highly oxidized state.

In figure 2, the times of escape of elements for Earth, Mars and Venus are plotted as a function of their atomic weights.[38] These calculations have been made on the assumption that the exospheric temperature of Mars is 1100°K,[4] and that of Venus is 2600°K.[5]

Venus, according to these considerations, must have also lost its hydrogen, while still retaining all the heavier elements. Mars loses its lighter elements at a still faster rate but the escape calculations suggest, however, that some atomic oxygen may have been retained on Mars till present time.

The curve for Jupiter is not shown in figure 2, because the time of escape, even for hydrogen, is of the order of 10^{200} years. We therefore expect that the atmosphere of Jupiter and of other giant planets would still have a composition similar to that of the solar atmosphere.

The atmosphere of the earth is composed of 78% nitrogen and 21% oxygen, with only traces of CO_2, H_2O and argon. However, in the atmosphere of another planet, which may have a similar molecular composition, the relative abundances of these molecules may be entirely different. On the earth the presence of free oxygen is probably a special circumstance which is connected with the presence of life. The CO_2 abundance in the atmosphere is limited by the reaction of the gas with the crust and the formation of carbonates, while the amount of water vapor is extremely sensitive to the temperature environment of the planet.

Therefore, although from the considerations of gravitational escape of gases we may expect the same moleculer composition for Mars and Venus, no predictions can be made regarding the relative abundances of these gases in the atmospheres of these planets.

The earth's atmosphere is "mixed" up to the altitude of 90 km, and therefore on an average the composition does not vary as a function of height. The mean molecular weight in approximately 29, and the total pressure at the surface is 10^6 dynes/cm^2. The number density at the ground is 2×10^{19} molecules/cm^3.

For an isothermal atmosphere in hydrostatic equilibrium, the variation of density with height

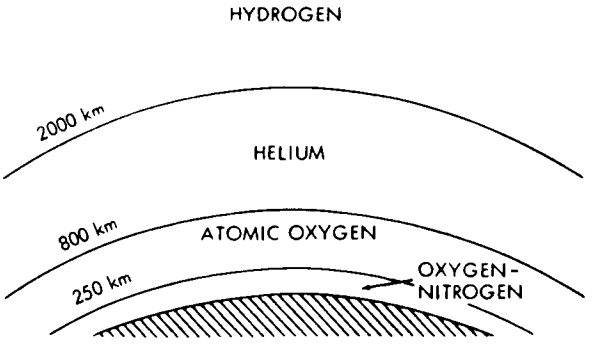


FIGURE 3.—Composition of the upper atmosphere of earth.

is given by the formula

$$\rho = \rho^0 \exp[-(h-h_0)mg/kT],$$

where ρ and ρ_0 are the densities at height h and h_0, m is the average molecular weight per particle, T the temperature between h and h_0, g the acceleration of gravity, and k Boltzmann's constant.

The quantity kT/mg is the scale height H of the isothermal atmosphere; at an altitude of one scale height the density is reduced by a factor of e. For a mean temperature of 250°K in the first 80 km of the atmosphere, the atmospheric scale height is approximately 8 km.

Above 120 km the diffusive separation sets in and the pressure of each constituent varies in accordance with the scale height calculated for its own molecular weight. Figure 3 shows the change in composition in the earth's atmosphere with altitude as deduced from the recent experimental results obtained by means of satellites.

For Mars, Venus, and Jupiter, we expect the lower atmospheres to be mixed also while the upper atmospheres of these planets will be in diffusive equilibrium; the major constituent of the atmosphere at the extremity will, in each case, be hydrogen.

Temperature

The temperature of a planet depends on its distance from the sun. In the case of the earth, the solar radiation flux, reaching the top of the earth's atmosphere has a value of about 1.4×10^6 ergs/cm²-sec. Part of this radiation (39%) is immediately "reflected" back to space by clouds and the atmosphere and does not play any role in the energy balance of the planet. Most of the remaining 61% of the solar radiations, mainly composed of the visible part of the spectrum, penetrates down to the ground and heats the surface to a certain temperature denoted as T_e, the effective temperature of the planet. In the case of rapidly rotating planets (e.g., Earth, Mars, Jupiter),

$$T_e^4 = (SC/4\sigma)(1-A),$$

where SC is the solar constant or the radiation flux received at the top of the atmosphere, and A is the albedo of the planet over the whole solar spectrum. For the values of sc and albedo given above, the T_e for earth is 245°K.

TABLE 1.—*Greenhouse Effect on Planets*

Planet	T_e (°K)	T_G (°K)	T (Greenhouse) (°K)	Optical thickness of atmosphere in the infrared
Mars	209	230	21	0.6
Earth	245	300	55	1.7
Venus	235	600	365	55.4
Jupiter	105	130	25	2.4

The surface radiating at this relatively low temperature emits primarily in the far infrared. A large fraction of the radiation emitted by the surface is immediately absorbed by molecules of CO_2 and H_2O present in the lower layers of the earth's atmosphere. A part of this absorbed energy in the infrared is returned to the surface and provides extensive heating of the ground, raising the surface temperature to the observed value of approximately 290°K.

This additional heating of the surface (about 45°K) by the return of infrared radiation from the atmosphere is referred to as the "greenhouse effect" and is an essential element in understanding the radiation budget of the planet.

The magnitude of the "greenhouse effect" caused by a planetary atmosphere can be estimated by approximating the solution to the radiative transfer equation. The ground temperature T_G is then given by the equation

$$T_G^4 = T_e^4(1+\tfrac{3}{4}\tau) \qquad (2)$$

where T_e is the effective black body temperature of the planet and τ is the total optical thickness of the atmosphere in the infrared. This solution is based on the assumption that the atmosphere is in radiative equilibrium. In the case of earth for the observed value of average ground temperature of 300°K, the total optical thickness of the atmosphere from equation 2 is about 1.7 which corresponds to the infrared absorption of about 82%. It is fairly well known that the amount of water vapor and the CO_2 in the earth's atmosphere do actually have an average opacity of about 80% over the entire infrared spectrum.

In table 1 the values of the effective black body

temperatures are given for other planets and also the most reliable values of the observed surface temperatures. In the case of Jupiter, T_G is the cloud top temperature and for that reason the factor $(1+\frac{3}{4}\tau)$ in equation 2 becomes $(\frac{1}{2}+\frac{3}{4}\tau)$. The differences give the magnitude of the greenhouse effect for each planet which in turn provide an estimate of the optical thickness of the atmosphere in the infrared.

On Mars the greenhouse effect is very small which gives an optical thickness of the atmosphere of only 0.6, indicating that the atmosphere is optically thin; therefore, we do not expect excessively large amounts of polyatomic gases like H_2O, CO_2, etc., which have intense absorption bands in the far infrared. On the other hand, for Venus, the magnitude of the greenhouse effect is extremely large compared to the earth or Mars, and the atmosphere must be almost completely opaque to the far infrared radiation from the planet. Large amounts of polyatomic gases should be abundant on Venus.

The atmosphere of Jupiter above the clouds should have an optical thickness of approximately 2.4 in order to explain the cloud top temperature of 130°K. Therefore we expect polyatomic gases, which absorb in the far infrared, at wavelengths greater than 20 μ where most of the planetary radiation will be confined.

With these considerations in mind, we now turn to the observed properties of the atmospheres of these planets.

MARS

The reliable physical data for Mars can be summarized as follows:

Mean distance from sun	1.52 AU
Mean equatorial diameter	6790 km
Length of day	1.0012 earth day
Length of year	1.8808 earth year
Mass	0.1078 (earth = 1)
Mean density	3.90 gm/cm^3
Gravity	377 cm/sec^2
Total integrated albedo	0.26
Effective black body temperature	209°K

Composition and Surface Pressure

The only gases that have been, up to now, spectroscopically detected in the atmosphere of Mars are carbon dioxide and water vapor. Their abundances have been estimated by several authors, but generally accepted values of the amounts of these gases in the atmosphere of Mars are not yet available. The best estimates, at the present time, vary within the following ranges:

$$CO_2 = 70 \pm 30 \text{ m-atm.}^{39}$$

Precipitable water vapor $= 10^{-3} - 10^{-2}$ gm/cm^2,[6,7] where m-atm is the thickness of a homogenous atmosphere in meters at normal temperature and pressure, 0°C and 760 mm.

The search for the presence of other gases in the atmosphere of Mars has failed to give positive results. The upper limits on their possible abundances are presented in table 2.

Earlier photometric studies by de Vaucouleurs[8] and polarimetric measurements by Dollfus[9] had given consistent estimates of the surface pressure on Mars as 85 ± 10 mb. Recently, however, from the analysis of the pressure-dependent absorption bands of CO_2 observed in the Martian atmosphere,[40] it has been suggested that the total pressure at the surface of Mars may be as low as 20 ± 10 mb. If this value is taken to be correct, then CO_2 will make up for about 25% of the total atmosphere.

The probable composition of the Martian atmosphere is presented in table 3. This atmosphere has a mean molecular weight of 32. With a surface pressure of 20 mb and an approximate temperature of 210°K, the surface density would

TABLE 2.—*Trace Constituents on Mars*

Gas	Upper limit of abundance (m-atm)
O_2	<2.4
N_2O, NO_2 or N_2O_4	<2.0
NH_3	<0.2
CH_4	<0.1
O_3	$<5\times10^{-4}$
SO_2	$<3\times10^{-5}$

TABLE 3.—*Probable Composition of Martian Atmosphere*

Gas	Volume (percentage)
N_2	72
CO_2	25
A	2
O_2	<0.5
H_2O	very small

be approximately 5×10^{17} particles/cm^3 which, for an isothermal atmosphere, should decrease exponentially with altitude with a scale height of 14.5 km The vertical distribution of pressure for this model is shown in figure 4. Results of a similar computation for earth, assuming an isothermal atmosphere at 250°K, are also shown in figure 4. It is interesting to note that although the pressure at the surface of Mars is about $\frac{1}{50}$ of the earth's at an altitude of approximately 50 km the two atmospheres have the same pressure, and above this height the pressure in the Martian atmosphere is greater than the pressure at the corresponding height in the earth's atmosphere. Because the decrease of density with height of Mars is almost two times slower than in the earth, the levels of the ionosphere and thermosphere on Mars would be much higher than on the earth.

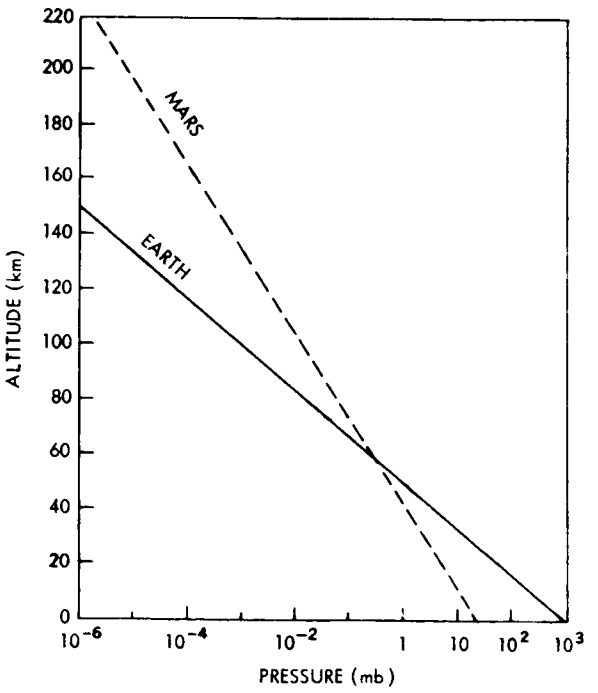


FIGURE 4.—Pressure vs. altitude for Earth and Mars. (Martian surface pressure about 20 mb, atmospheric scale height = 14.5 km.)

Temperature

Planetary temperatures are usually estimated by measuring the infrared radiation emitted by the planet. A large part of the infrared spectrum is, however, absorbed by the water vapor and CO_2 present in the earth's atmosphere. Ground based observations of the planets in the infrared are therefore confined to the 8–12 μ region where the earth's atmosphere is relatively transparent. This spectral region in the infrared is known as the atmospheric "window".

Extensive temperature measurements of Mars by infrared radiometry in this window of the terrestrial atmosphere have been made since 1926, and the results to date are quite consistent. Since the Martian atmosphere should also be largely transparent in the 8–12 μ window, it is assumed that the radiation intensity measurements in this wavelength region refer to the surface of the planet.

Some of the earlier investigators were also able to scan the planetary disc latitudinally and longitudinally so as to obtain the diurnal and seasonal variation of temperature as a function of latitude. The up-to-date information can be summarized as follows:

Maximum temperature at equator	~300°K
Mean amplitude of diurnal variation	
Noon to sunset	~ 60°K
On earth in desert	~ 30°K
Night side temperature cannot be measured but can probably be estimated at the equator	~200°K
Day side temperature at poles	~220°K
Mean temperature of day side	~260°K
Mean temperature of whole planet	~230°K

Surface Features

As seen by telescope the outstanding features of Mars are (1) the dark areas (maria), (2) the

general reddish-orange background ("deserts"), and (3) the polar caps.

The nature of the dark maria is not very well understood because, apart from showing fairly regular seasonal changes, they are also subject to erratic variations which make the hypothesis of earth-type vegetation somewhat doubtful. Moreover, the infrared spectrum does not have characteristic absorption bands of chlorophyll. The presence of organic material in these regions has, however, been reported by Sinton[10] who detected C–H vibration absorptions in the 3.5 μ region which were absent in the reflection spectrum of the desert areas.

The probability of the existence of life on Mars is controversial and beyond the scope of this review; but it will be desirable to see the observations of C–H bands on Mars repeated and probably substantiated by infrared spectroscopic measurement of the terrestrial "dark and bright" areas from an earth satellite.

The large orange areas of Mars are responsible for the reddish color of the planet seen by the naked eye. According to Kuiper[11] they are composed of felsitic rhyolite.

The polar caps are rather reliably known to be a layer of H_2O frost a few centimeters thick deposited on the surface. These caps are observed to expand to lower latitudes during fall and winter and to recede to high latitudes during spring and summer. The observed evaporation of the polar caps in summer and the almost immediate transport of the water vapor towards the winter pole across the equator is a special meteorological phenomenon peculiar to Mars. In the earth's atmosphere the hemisphere to hemisphere mixing is extremely small. The average speed of this "humidity wave" across the planet has been estimated to be about 45 km/day. This pole to pole circulation can perhaps only be understood by the greater length of the Martian year and the increased temperature differences between the cold and warm poles. In the summer the pole is the warmest region on the planet.

Blue Haze Layer

Photographs of Mars taken through blue filters indicate no surface features, and it is therefore believed to be covered with a haze layer known as "blue haze", absorbing at these wave lengths. Sometimes during the opposition, however, it clears over certain regions, and surface features become observable in the blue. These "blue clearings" last usually a few days. Conflicting arguments have been advanced to explain the nature of the blue haze, but an explanation which satisfies all the observed features is yet to be found.

Clouds

Several types of thin clouds have been frequently observed on Mars. They can be divided into three types:

Blue clouds are visible only in blue light. They appear in patches near the poles and the terminator. Polarization measurements indicate their particle size to be approximately 0.1 μ. They probably occur at altitudes less than 100 km. According to Goody,[12] they may be composed of ice crystals formed on the nuclei fed by the haze layer, while others believe them to be of the same material as the blue haze.

White clouds are visible both in blue and yellow light. The polarization studies of these clouds suggest their nature to be the same as ice crystal clouds of a particle size of about 1 μ. Being composed of larger particles, they probably lie at lower altitudes than the blue clouds.

Yellow clouds visible only in yellow, are very rare and variable in size. They have been seen drifting several hundred miles across the planet at a velocity of 60 km/hr. According to Goody,[12] they are composed of the same blue absorbing material as the blue haze. Hess[13] estimates their height to be approximately 5 or 6 km.

Vertical Distribution of Temperature

The atmosphere of Mars is optically thin in the infrared, and the probable absorbing gases are CO_2 and H_2O, which are also present in the earth's atmosphere. The vertical temperature profile in the atmosphere of Mars, therefore, cannot be determined by observation from the surface of the earth. Only theoretical estimates exist for the vertical temperature structure of Mars. The most recent computations of this nature are

by Arking, Ohring, and Chamberlain (references 14, 15, and 4, respectively).

Arking has calculated a model atmosphere for Mars, allowing for convection and using the exact equation of radiative transfer for frequency independent absorption. A total optical thickness of 0.5 was chosen to obtain a surface temperature of 235°K, consistent with the observed mean surface temperature. Assuming an effective black body temperature of 217°K, an adiabatic gradient of −3.7°K/km in the convection zone and an exponential dependence of absorption on altitude with a scale height of 17 km, the temperature profile shown in figure 4 (insert) is obtained. The convection zone is found to extend up to 8 km.

Ohring also has recently investigated the vertical temperature profile for a model Martian atmosphere containing 2% CO_2, 98% N_2 and no water vapor. The ground temperature was assumed to be 230°K and the transfer of radiation in the atmosphere was calculated for frequency dependent absorption by the CO_2. The tropopause in this case was found to be at 9 km at a temperature of 196°K. In the stratosphere the temperature keeps on decreasing to a value of as low as 90°K at an altitude of 42 km, where the total pressure is 2.5 mb.

Now that the mixing ratio of CO_2 in the Martian atmosphere is known to be much higher, new calculations for a "non-grey" atmosphere are desirable.

The above mentioned results of the temperature distributions in the Martian atmosphere do not take into account possible heating of the lower atmosphere by direct absorption of solar radiation in the ultraviolet. In the case of the terrestial atmosphere, the ozone heating produces a temperature maxium at 50 km, but on Mars, with much less oxygen compared to earth, ozone would be confined to lower layers of the atmosphere. The solar radiation in the region of 2500 A, responsible for ozone dissociation and atmospheric heating, will therefore penetrate to a much lower depth in the Martian atmosphere. Due to a higher atmospheric density at this level (compared to the density in the earth's atmosphere at 50 km) and because of the lower intensity of the solar radiation at the distance of Mars, the heating rates will probably not be as important as in the ozonosphere on earth. Still they may affect the temperature gradients substantially and limit the extent of the convection zone.

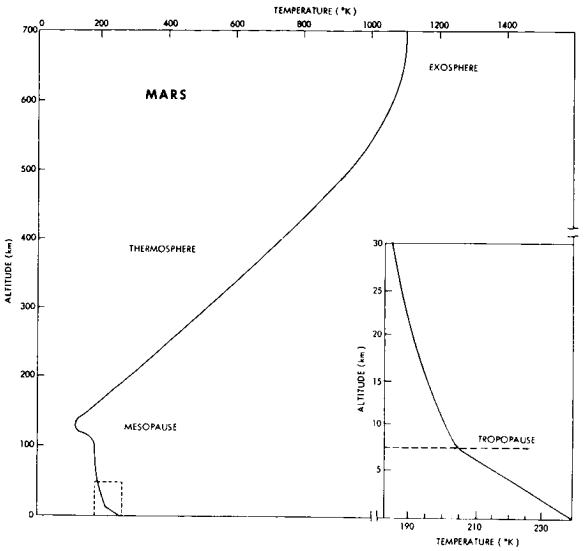


FIGURE 5.—Vertical temperature structure of the Martian atmosphere (from references 4 and 14).

Assuming an atmospheric composition of 98% N_2 and 2% CO_2, and a Martian stratospheric temperature of 134°K,[12] Chamberlain[4] recently computed the mesospheric cooling and thermospheric heating for Mars. With the assumed model atmosphere and from the considerations of CO_2 dissociation into CO and O and the consequent CO cooling at the mesopause, Chamberlain deduces the height of the mesopause as about 130 km at a temperature of 76°K. The considerable CO cooling at the mesopause level acts as "an effective thermostat, keeping the temperature at the exospheric or escape level (1500 km) from exceeding 1100°K."

Figure 5 shows the vertical distribution of temperature in the Martian atmosphere. The temperatures up to an altitude of 100 km are based on the results obtained by Arking. Above this altitude, the upper atmospheric model developed by Chamberlain has been used.

VENUS

Venus is our nearest planetary neighbor and, after the sun and moon, the brightest object in

TABLE 4.—*Physical Constants of Venus and Earth*

Planet	Mass	Radius	Distance (AU)	Density (gm/cm^2)	Albedo	T_e (°K)	g (cm/sec^2)
Earth	1	1	1	5.5	0.39	245	980
Venus	0.81	0.97	0.72	5.0	0.73	235	842

the sky. It has therefore attracted the attention of man since the beginning of civilization. Despite the great interest, however, very little is known about the atmosphere of this planet, especially when compared with the information we have about Mars. The main reason for this deficiency is that Venus is covered with a layer of white clouds, and the surface remains permanently invisible. Observations have, however, been made in the infrared and radio frequency regions, and new information regarding the composition and temperature distribution in the Venus atmosphere has been obtained in the last few years. These new data on the temperature and pressure at the various levels in the Venus atmosphere have forced a complete revision of ideas regarding the atmospheric structure of this planet. We shall try to summarize our present-day knowledge regarding the Venus atmosphere by first giving the physical constants of the planet (table 4) and then reviewing the current hypotheses regarding its structure.

Composition

From the analysis of the reflected solar spectrum, the only constituent of Venus atmosphere established beyond question in CO_2. Its abundance above the effective "reflecting level" of the sunlight has been estimated by several workers. Recent reinterpretations of the old spectra of Venus by Spinrad[16] give a CO_2/atmosphere ratio of only 5% by mass.

The scattering and polarizing properties of the clouds correspond to fine droplets of about 2 μ in diameter with the refractive index of water. However, Spinrad has carefully examined a high dispersion spectrogram of Venus taken by Adams and Dunham[17] at a time when there were apparently "breaks" in the clouds, and he has concluded that the mixing ratio of water vapor down to a pressure of 8 atmospheres was less than one part in 10^5.

At the same time, Dollfus[18] has identified water vapor absorption in the upper atmosphere of Venus, and obtains a mixing ratio of 10^{-4} or 10^{-5} by mass, depending upon the cloud top pressure.

From the observed equality in the radar reflectivity of Venus at 68 and 12.5 cm,[19] Thaddeus[41] has recently calculated an upper limit of 3 gm/cm^2 to the total amount of water vapor which could be present in the atmosphere of Venus. From these results the atmosphere of Venus seems to be extremely dry, but a generally accepted value for the amount of water vapor present is not yet available.

Sinton[20] has presented evidence for the existence of CO above the reflecting level of two micron photons. However, Kuiper[21] has not detected this gas in his recent analysis of high resolution spectra of Venus.

Urey[22] has noted that as much as 80 m-atm of oxygen could exist above the clouds without having been detected. Also according to Urey, CH_4 and NH_3 cannot be important constituents of the atmosphere in the presence of CO_2, and the oxides of nitrogen cannot be expected in any planetary atmosphere because they are unstable against decomposition into N_2 and O_2. It seems likely that the bulk of the atmosphere is composed of N_2. Rare gases may also be present as minor constituents.

The likely composition of the atmosphere of Venus is therefore as follows: N_2, 95%; CO_2, 5%; and H_2O less than 10^{-3}% by mass.

Pressure

There is conflicting evidence concerning the pressure at the cloud top level. A difference in the polarization of red and green light reflected

from the cloud top has been interpreted by Dollfus[23] as the result of molecular scattering in an atmosphere 800 m. thick at STP, which corresponds to a cloud top pressure of 90 mb. Sagan[24] has analyzed Spinrad's results on the temperature-pressure combinations indicated by CO_2 lines on different plates, and he concludes that the cloud top pressure on the illuminated side lies between 0.53 and 0.83 atm. This is consistent with Kaplan's[25] discussion of the pressure broadening of the 1.6 μ CO_2 bands.

From other considerations Sagan also derives a cloud top pressure of 90 mb for the night side of venus with an uncertainty of a factor of 3.

Estimates for the pressure at the surface of Venus range from 7 to 200 atm. These will be discussed in more detail in a later section.

Temperature

Cloud Top

Sinton and Strong[26] have repeatedly measured the planetary emission from the cloud top in the 8–12 Ω region and find a temperature value of 235±10°K. This temperature probably refers to the cloud top level of the Venus atmosphere. Recent results from Mariner II seem to confirm this conclusion.[27]

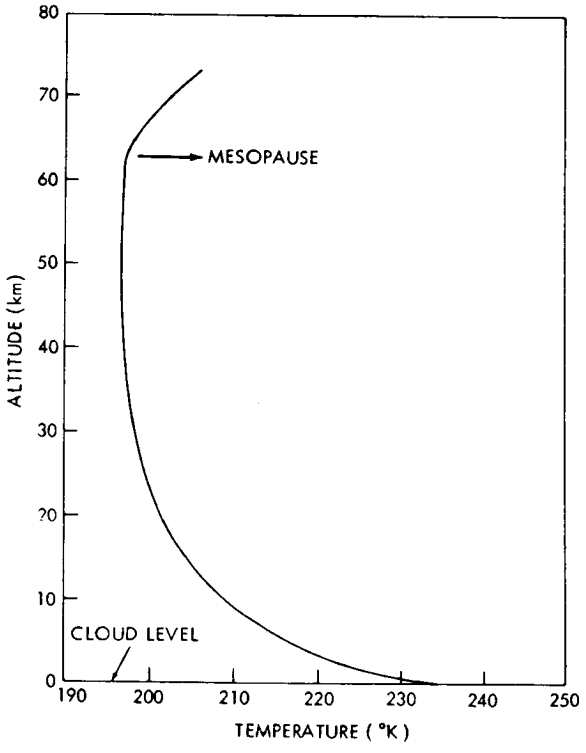


FIGURE 6.—Temperature profile in the Venus atmosphere above the clouds (from reference 28).

From this value of temperature it is possible to construct a model for the atmosphere of Venus above the clouds. Figure 6 shows a temperature profile computed by Rasool[28] on the assumption that the atmosphere is in radiative equilbrium above the clouds.

The temperature should reach an asymptotic value of 197°K above an altitude of about 20 km. The mesopause will probably occur above 60 km, and then the temperature will rise in the thermosphere because of heating by photodissociation and photoionization reactions. Lack of knowledge of the composition prevents predicting whether heating should occur below the mesopause by such mechanisms as ultraviolet absorption by ozone.

Surface

Determinations of the surface temperature of Venus by optical astronomy have been prevented by the cloud cover. However, the small amount of thermal radiation emitted by the planet in the centimeter wavelength region should penetrate through the clouds without significant attenuation, and therefore it can be used to determine the temperature of the surface of the planet.

First attempts to measure radiation from Venus in the microwave region were made in 1956 by the radiotelescope of the Naval Research Laboratory. The temperature inferred from the measured radiation intensity, was, however, unexpectedly high—of the order of 600°K, which is certainly too hot to support any imaginable form of life. Repeated measurements in following years have confirmed this result.[29]

The observed temperatures are plotted as a function of wavelength in figure 7. At longer wavelengths the brightness temperature is observed to increase when the Venus-sun-earth angle becomes appreciable, indicating that the illuminated hemisphere is significantly hotter than the dark side. (The 10 cm. brightness temperature may exceed 700°K on the bright side.) Three

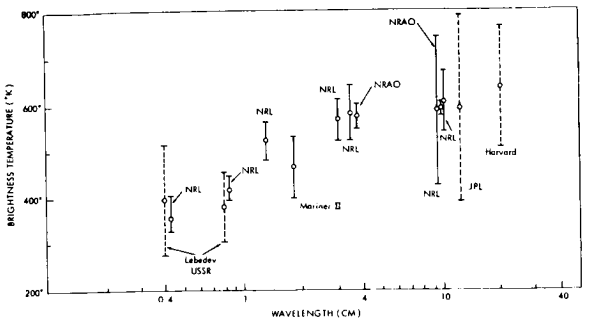


FIGURE 7.—Observed microwave brightness temperatures of Venus at different wavelengths.

models of Venus have been considerably discussed in connection with these measurements.

Greenhouse Model

First is the greenhouse model originally discussed by Sagan[30] in which it is assumed that the 600°K radio-brightness temperature actually exists at the surface of the planet and is maintained by a very effective greenhouse effect. The solar radiation in the visible penetrates up to the surface; the planet, thereby being heated up to a temperature T_e, emits in the infrared; but due to the presence of triatomic molecules like CO_2 and H_2O, which have strong absorption bands in the infrared region, most of the radiation remains trapped in the atmosphere and heats up the surface.

Jastrow and Rasool[31] have recently pointed out that in such a model an extremely high infrared opacity of the atmosphere (optical thickness of 42, meaning a transmission of only 10^{-18}) will be required in order to obtain a Venus ground temperature of 600°K.

It is very difficult to understand how the lower Venus atmosphere can be so opaque in the infrared and still be transparent in the visible. Tri-atomic molecules are good infrared absorbers, but they have "windows" at certain wavelengths which must be blocked by other molecules if the opacity is to be significant. Jastrow and Rasool have also shown that the combination of CO_2 and small amounts of H_2O at pressure of the order of 3 atm. still contains significant windows.

However, the situation alters if the atmospheric pressure at ground level is many tens or hundreds of atmospheres. Then pressure broadening may close all the windows. Plass and Stull[32] have shown that the optical thickness approaches 5 if the ground pressure is about 30 atms. and the pressure broadening of CO_2 alone is considered. At the high temperatures on Venus there may be many trace constituents in the atmosphere that significantly increase the opacity.

Aeolosphere

The second model, called the aeolosphere model, is due to Öpik[33] who suggests that the blanketing must be due to dust, wind friction of the surface being the main source of energy for the high temperature. The dust is probably made of calcium and magnesium carbonates and the atmosphere is mainly composed of CO_2 and N_2. No water vapor is necessarily present in the atmosphere. Due to the blanket of dust there is no sunlight penetrating to the surface. This model predicts that the microwave brightness temperatures of Venus should not change with the phase of Venus, which is apparently contradicted by the observations.

Ionospheric Emission

A third model invokes a dense ionsphere of Venus which is assumed to emit the observed microwave radiation. This model predicts that the radio observations should show a "brightening" at the limb of Venus, whereas the observations made by Mariner II have recently indicated "limb darkening".[34]

JUPITER

Relatively little is known about the structure of the atmosphere of Jupiter.

Being five times farther removed from the sun than the earth, but having a comparable albedo of 0.47, the effective temperature of the planet is only 105°K. The high surface gravity and low temperature together suggest that the gravitational escape of gases from the planet should be extremely inefficient and the chemical composition of the atmosphere should still be primitive, containing large quantities of hydrogen and helium.

TABLE 5.—*The Approximate Composition of the Jovian Atmosphere*

Gas	Volume (Percent)
H_2	60
He	36
Ne	3
CH_4 and NH_3	<1

Composition

The only gases detected spectroscopically in the atmosphere of Jupiter are NH_3, CH_4 and hydrogen. The abundances of these gases have been estimated by various authors, and table 5 gives the approximate composition of the atmosphere of Jupiter.[35]

Helium is spectroscopically not observable from the earth and has been added due to its cosmic abundance and to account for the estimates of the cloud top pressure of 3 atms.

Table 5 indicates that the Jovian atmosphere may not be as rich in hydrogen as expected from the discussion of gravitational escape.

Measurements of the scale height of the atmosphere of Jupiter have been made by observing the occulatation of σ-Arietis by Jupiter. These observations yield a scale height of about 8 km. For a mean atmospheric temperature of 120°K, the corresponding molecular weight would be approximately 4. This requires more helium than table 5 indicates. However, according to Gallet[36] the relative abundance of H and He given in table 5 could also account for a mean molecular weight of 4 if the Jovian atmosphere had a substantial amount of neon.

Temperature

The calculated black body temperature of Jupiter for a visual albedo of 0.47 is 105°K. Spectroscopic measurements in the infrared indicate a temperature of 130°K (reference 37) which probably refers to the cloud top level.

Gallet[36] has made extensive theoretical analyses of the atmospheric structure of Jupiter below the clouds. The visible clouds have been assumed to be made of NH_3 crystals, which condense at the temperature and density of the cloud top. With allowance for the heat released in this condensation, Gallet calculates a relatively slow increase of temperature going down from the clouds to the surface and thus arrives at high densities and pressures at low temperatures. Gallet has also suggested that the observed cloud layer may be NH_3 of thickness 50 km, below which there is another deck of clouds composed of H_2O, Between these two clouds is a region where NH_3 rain storms can occur. In such a model the surface of Jupiter would be a few hundred km below the clouds, at a temperature of about 1000°K and a pressure of several thousand atmospheres.

REFERENCES

1. ALLER, L. H., "The Abundance of the Elements," New York: Interscience Publishers, 1961.
2. SPITZER, L., Jr., "The Terrestrial Atmosphere above 300 Km," in "The Atmospheres of the Earth and Planets," (G. P. Kuiper, ed.): 211–247, Chicago: University of Chicago Press, 1952.
3. MACDONALD, G.J.F., "The Escape of Helium from the Earth's Atmosphere," *Revs. Geophys.* (in press, 1963).
4. CHAMBERLAIN, J. W., "Upper Atmospheres of the Planets," *Astrophys. J.* **136**(2): 582–593, September 1962.
5. WALKER, J. C. G., and JASTROW, R., "The Thermospheres of the Planets," *Amer. Geophys. Union Trans.* **44**(1): 85, March, 1963.
6. SPINRAD, H., MÜNCH, G., and KAPLAN, L. D., "The Detection of Water Vapor on Mars," *Astrophys. J.* **137**(4): 1319–1325, May 1963.
7. DOLLFUS, A., "A Measurement of the Quantity of Water Vapour Contained in the Atmosphere of the Planet Mars," (In French), *C. R. Acad. Sci.* **256**(14): 3009–3011, April 1963.
8. VAUCOULEURS, G. H. DE, "Physics of the Planet Mars," London: Faber and Faber, 1954.
9. DOLLFUS, A., "Études des planètes par la polarisation de leur lumiere," *Ann. de Astrophisique* Supp. no. 4, 1957.
10. SINTON, W. M., "Further Evidence of Vegetation on Mars," *Science* **130**(3384): 1234–1237, November 6, 1959.
11. KUIPER, G. P., "Planetary Atmospheres and Their Origins," in "The Atmospheres of the Earth and Planets," (G. P. Kuiper, ed.): 306–405, Chicago: University of Chicago Press, 1952.
12. GOODY, R. M., "The Atmosphere of Mars," Weather **12**(1): 3–15, January 1957.
13. HESS, S. L., "Mars as an Astronautical Objective," in *Advance in Space Science and Technology*, v. 3 New York: Academic Press, 1961, pp. 151–193.

14. ARKING, A., "Non-Grey Convective Planetary Atmospheres," Soc. Roy. Sci. Liège Mem. 7 (spec. no.): 180–189, 1963.
15. OHRING, G., "A Theoretical Estimate of the Average Vertical Distribution of Temperature in the Martian Atmosphere," *Icarus* **1**(4): 328–333, January 1963.
16. SPINRAD, H., "Spectroscopic Temperature and Pressure Measurements in the Venus Atmosphere, *Astron. Soc. Pacific Publ.* **74:** 187–201, June 1962.
17. ADAMS, W. S., and DUNHAM, T., JR., "Absorption Bands in the Infra-Red Spectrum of Venus," *Astron. Soc. Pacific Pub.* **44**(260): 243–245, August 1932.
18. DOLLFUS, A., "Observations of Water Vapour on the Planet Venus," (In French), *C. R. Acad. Sci.* **256**(15): 3250–3253, April 8, 1963.
19. MUHLEMAN, D. O., "The Electrical Characteristics of the Atmosphere and Surface of Venus from Radar Observations," *Icarus* **1**(5/6): 401–411, April 1963.
20. SINTON, W. M., "Infrared Observations of Venus," Soc. Roy. Sci. Liège Mem. 7 (spec. no.): 300–310, 1963.
21. KUIPER, G. P., "Infrared Spectra of Planets and Cool Stars," Onzieme Colloque International d'Astrophysique, La Physique des Planètes, Universitè de Liège, 1963.
22. UREY, H. C., "The Atmospheres of the Planets," in "Handbuch der Physik," (S. Flügge, ed.), **52:** 363–418, Berlin: Springer Verlag, 1959.
23. DOLLFUS, A., "Polarization Studies of Planets," in "The Solar System III, Planets and Satellites," (G. P. Kuiper and B. M. Middlehurst, eds.): 343–399, Chicago: The University of Chicago Press, 1961.
24. SAGAN, C., "Structure of the Lower Atmosphere of Venus," *Icarus* **1**(2): 151–169, September 1962.
25. KAPLAN, L. D., "A New Interpretation of the Structure and CO_2 Content of the Venus Atmosphere," *Planet Space Sci.* **8**(1): 23–29, October 1961.
26. SINTON, W. M., and STRONG, J., "Radiometric Observations of Venus," *Astrophys. J.* **131**(2): 470–490, March 1960.
27. CHASE, S. C., KAPLAN, L. D., and NEUGEBAUER, G., "Mariner II: Preliminary Reports on Measurements of Venus: Infrared Radiometer," *Science* **139:** 907–908, March 8, 1963.
28. RASOOL, S. I., "Structure of Planetary Atmospheres," *Amer. Inst. Aeronaut. Astronaut. J.* **1**(1): 6–19, January 1963.
29. MAYER, C. H., "Radio Emission of the Moon and Planets," in "The Solar System, Vol. III, Planets and Satellites," (G. P. Kuiper, and B. M. Middlehurst, eds.): 442–472, Chicago: University of Chicago Press, 1962.
30. SAGAN, C., "The Radiation Balance of Venus," JPL Tech. Rep. No. 32–34 (NASA Contract No. NASw-6), Pasadena: California Institute of Technology, September 15, 1960.
31. JASTROW, R., and RASOOL, S. I., "Radiative Transfer Transfer in the Atmospheres of Venus and Mars," in Space Research III, ed. by W. Priester, Amsterdam: North Holland Publishing Co., 1963.
32. PLASS, G. N., and STULL, V. R., "Carbon Dioxide Absorption for Path Lengths Applicable to the Atmosphere of Venus," *J. Geophys. Res.* **68**(5): 1355–1363, March 1, 1963.
33. Öpik, E. J., "The Aeolosphere and Atmosphere of Venus," *J. Geophys. Res.* **66**(9): 2807–2819, September 1961.
34. BARATH, F. T., BARRETT, A. H., COPELAND, J., JONES, D. E., and LILLEY, A. E., "Mariner II: Preliminary Reports on Measurements of Venus: Microwave Radiometers," *Science* **139:** 908–909, March 8, 1963.
35. SPINRAD, H., and TRAFTON, L. M., "High Dispersion Spectra of the Outer Planets. I. Jupiter in the Visual and Red," *Icarus* **2**(1): 19–28, June 1963.
36. GALLET, R., "The Structure of the Cloud Layers and the Deep Atmosphere of Jupiter," in "Proc. Conf. on the Planet Jupiter," (H. J. Smith, ed.), (In press).
37. MURRAY, B. C., and WILDEY, R. L., "Stellar and Planetary Observations at 10μ," *Astrophys. J.* **137**(2): 692–693, February 15, 1963.
38. STEWART, Private communication, 1963.
39. L. D. KAPLAN, Private communication, 1963.
40. L. D. KAPLAN, Private communication, 1963.
41. P. THADDEUS, Private communication, 1963.

BIBLIOGRAPHY

CAMERON, A. G. W., "Physics of the Planets," in *Space Physics* (In press, 1963).

HESS, S. L., "Mars as an Astronautical Objective," in "Advances in Space Science and Technology," v. 3 (F. I. Ordway, ed.): 151–193, New York: Academic Press, 1961.

JASTROW, R., "Results of Experiments in Space," *J. Aerospace Sci.* **29,** 377 (1962).

KELLOGG, W. W., and SAGAN, C., "The Atmospheres of Mars and Venus," Washington, D. C.: Nat. Acad. Sci.-Nat. Res. Coun. Publ. 944, 1962.

NEWBURN, R. L., JR., "The Exploration of Mercury, the Asteroids, the Major Planets and Their Satellite Systems, and Pluto," in "Advances in Space Science and Technology," v. 3 (F. I. Ordway, ed.): 195–272, New York: Academic Press, 1961.

SAGAN, C., and KELLOGG, W. W., "The Terrestrial Planets," *Ann. Rev. Astron. Astrophys.* **1:** 235–266, 1963.

WILDT, R., SMITH, H. J., SALPETER, E. E., and CAMERON, A. G. W., "The Planet Jupiter," *Phys. Today,* **16**(5): 19–23, May 1963.

RESULTS OF MICROMETEORITE PENETRATION EXPERIMENT ON THE EXPLORER VII SATELLITE (1959 IOTA)

H. E. LAGOW AND L. SECRETAN

Goddard Space Flight Center

The results of a micrometeoroid penetration experiment aboard satellite 1959 IOTA (Explorer VII) are presented. The sensors and their arrangement aboard the satellite are described, and the telemetry record obtained during flight is analyzed critically. It is concluded that one penetration through one cell occurred on the 16th day, and that it was caused by a particle approximately 10 microns in diameter.

INTRODUCTION

Explorer VII (1959ι) was launched on October 13, 1959 by a Juno II rocket from the Atlantic Missile Range, Cape Canaveral, Florida. A part of the instrumentation of this satellite was designed to measure micrometeorite penetration or molecular sputtering by utilizing photo-conducting CdS cells. This publication presents a description of these experiments and of the results obtained from them.

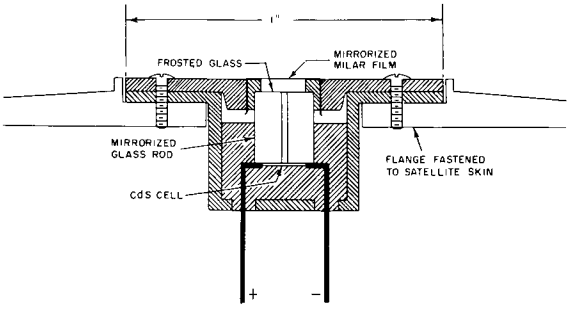


FIGURE 1.—Cross section of a CdS cell such as the ones used on Explorer VII.

Figure 1 shows a cross section of a CdS cell, which has been described elsewhere.[1] On 1959 Iota, three CdS cells were mounted on a magnesium plate located on the satellite's equator in thermal contact with the battery brackets through solid copper bars. The three CdS cells, facing outward perpendicular to the satellite's spin axis, were identical in design and in effective area (18 mm^2), but had covers differing from each other in the following manner:

Cell 1 was covered with $\frac{1}{4}$-mil Mylar film coated with an opaque layer of aluminum, approximately 1000 angstroms thick, on the front side only. This cell would be sensitive to both sputtering and penetration.

Cell 2 was covered in the same manner on *both* sides. This cell would be sensitive primarily to penetration.

Cell 3 had a 1-mil aluminum cap with a perforation 0.0075 mm^2 in area for calibration purposes.

Cell 2 was in addition, equipped with a 5-kilohm bead thermistor for temperature measurements of these sensors.

The resistances of these three sensors, the thermistor, and two calibration resistors were multiplexed into a resistance-controlled oscillator whose nominal frequency range was from 680 to 780 cps. The subcarrier frequency phase-modulated a 108.5-Mc tracking transmitter.[2] The

multiplexed channels were allocated as follows:

Channel	*Sensor*
1	Calibration 700 Ω resistor
2	CdS Cell 1
3	CdS Cell 2
4	CdS Cell 3 (perforated cap for calibration)
5	Temperature sensor
6	Calibration 20K resistor

CALIBRATION

The CdS cells were calibrated in terms of their response to the sun or to a xenon arc for various amounts of light input as regulated by calibrated holes through opaque covers. Rotating the CdS cells at 360 rpm in the light beam simulated the effect of the rotating satellite. Photometric measurements were taken in correlation with microscopic measurements of the area of the hole. The accuracy of these measurements was an order of magnitude better than the resolution expected from the CdS cells in their response to light stimuli. The change of response to sunlight as a function of the sun's angle with the normal to the cells was also determined. Finally, the CdS cells were calibrated for temperature changes within the expected temperature range ($-10°$ to $+60°$C; see reference 3).

RESULTS

The three CdS sensors and the thermistor performed normally and the data obtained from them are presented in the following paragraphs.

Cell 1, with a single aluminum coating, indicated a perforation of its cover before termination of the coasting period (between first and second-stage powered flight). The cause of this perforation, which occurred 540 seconds after lift-off and several telemetry frames after the ejection of the protective shroud, could not be determined. The hole is estimated to have been $0.16 \pm .04$ mm in diameter; its size (approximately 0.1 percent of the cell's total area) remained virtually constant through the lifetime of the experiment.

The calibration cell (Cell 3) responded as expected to sunlight until the end of battery life, and was used in the computation of the correction applied to the results obtained from Cell 2.

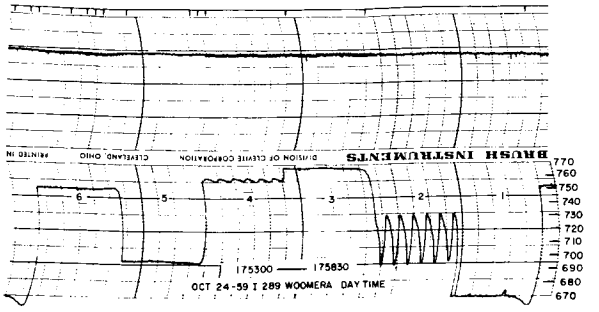


FIGURE 2.—Explorer VII telemetry record of a selected daylight pass on October 24, 1959.

Cell 2 registered darkness from the launching on October 13 until October 26, 1959 (see figure 2 for a selected daylight pass on October 24). On October 29 the telemetering record (figure 3) suggests that sunlight is entering the cell. Records for October 27 and 28 were not sufficiently clear to distinguish a deviation as small as that seen in figure 3. When this record is compared with figure 4, a nighttime record made on November 7, 1959, it is evident that the Cell 2 (Channel 3) *scale* reading is about 3 percent (of the 100-cps interval plotted) lower in daylight than in darkness for all times after October 29.

Further evidence of light penetration can be seen from the Cell 1 and Cell 3 records (Channels 2 and 4) in figure 3. They show characteristic sawtooth patterns corresponding to the spin rate of the satellite; the steeper slope in the sawtooth corresponds to the beginning of the light input to the cell, and the lesser slope to the dark phase of the spin cycle. The amplitude and shape of the

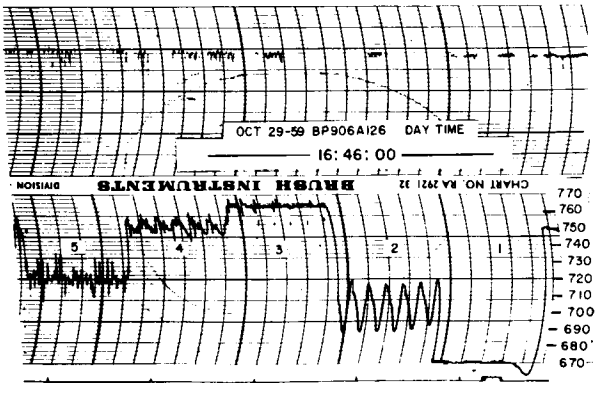


FIGURE 3.—A daytime pass on October 29, 1959.

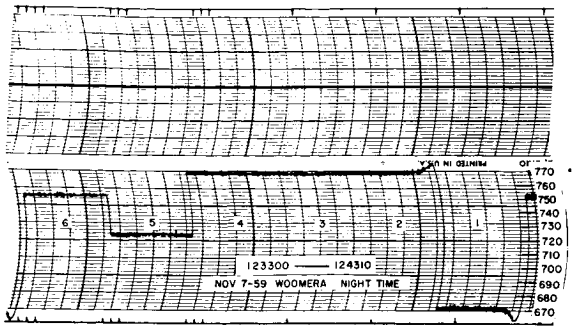


FIGURE 4.—A nighttime pass on November 7, 1959.

sawtooth signal are functions of the amount of light entering the cell.

In the case of Cell 2 which was located physically and connected electrically in sequence between Cells 1 and 3, it is expected that the sawtooth pattern of Cell 2 should follow the pattern of Cell 1 and blend into the pattern of Cell 3, since the switching time from cell to cell is extremely short. Careful observation of the October 29 record (figure 3) shows that a very small pattern does indeed seem to be present (Channel 3), marked by a repetition of ticks that fall in cycle with the sawtooth discussed above.

It is concluded that the 3 percent change in frequency, while being approximately at the limit of detectability, is definitely readable. In terms of ohmic resistance, it indicates a change from the

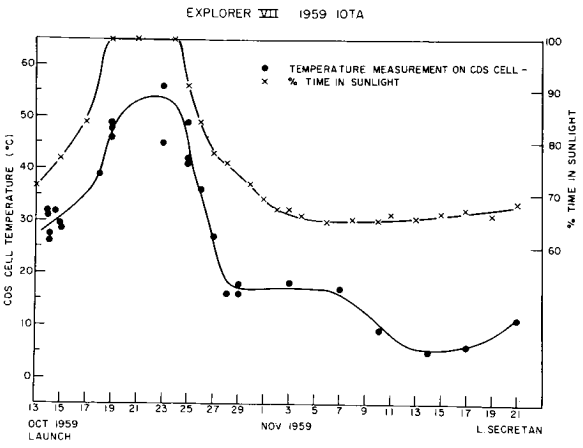


FIGURE 5.—Measured CdS cell temperatures and calculated percent times in sunlight for Explorer VII from October 13 (launch) to November 21, 1959.

100K dark reading to 60–70K, which on a rotating satellite and at normal light incidence, corresponds to an opening having an area of 10^{-4} mm^2. This apparent area of the opening must be corrected for the sun angle. Direct analysis of the signal from the calibration cell (Cell 3) shows that the actual area is twice the apparent area. Thus, the area of the opening in Cell 2 is 2×10^{-4} mm^2, corresponding to a hole diameter of 16 microns. Further corrections for temperature were not necessary because the temperature at this time was near room temperature (figure 5).

DISCUSSION

Satellite 1959ι exposed the three CdS cells to the space environment of micrometeorites, trapped radiation, and sputtering for 38 days of active life. Cell 1 was perforated during the launching and Cell 3 was relatively insensitive because its primary function was calibration. The telemetry signals from Cell 2 show:

1. That for the first 13 days, neither sunlight nor energetic particles caused a readable response.
2. That on the 16th day sunlight was entering the cell, causing a response that corresponded to a hole size of 2×10^{-4} mm^2 (about 16 microns in diameter).
3. That after the 16th day recordings made when the satellite was in darkness showed normal dark response.
4. That the hole size did not change by a readable amount from the 16th day to the 38th.

It is concluded that the response on Cell 2 after the 16th day was due to a micrometeorite penetration of the $\frac{1}{4}$-mil Mylar and opaque aluminum coatings on both faces.

The relation of the hole size to the pentrating particle has been defined in the recent work with the hypervelocity accelerator at Space Technology Laboratories, showing that a hole in $\frac{1}{4}$-mil Mylar is about 3 times the size of the penetrating projectile.[4] In the present case, then, the perforating particle would be 10 microns in diameter. Assuming a density of 1 to 3 gm/cm^3, the mass is between 5 and 15×10^{-10} gram. From the satellite microphone data[5] the influx rate for particles of mass 5 to 15×10^{-10} gram is between 6×10^{-2} and 8×10^{-3} impacts/m^2 per second. In this case

the area is 18 mm^2, the time 38 days, and the probability of one hit lies between 0.9 and 0.27.

It should be noted that there is a finite possibility that the hole was produced by molecular sputtering; however, this is improbable because it requires a latent hole in the inner coating.

ACKNOWLEDGMENTS

The authors wish to acknowledge the assistance of the following persons in performing this experiment. Dr. H. Shenker of the U. S. Naval Research Laboratory and Mr. M. Lubin of JEM Electronics for development of the CdS sensors. Mr. Olin King of Marshall Space Flight Center for the development of the telemetering muliplexer and transmitter. Mr. G. B. Robinson of Goddard Space Flight Center, for processing the telemeter tapes. Mr. W. M. Alexander for discussion of the results.

REFERENCES

1. LaGow, H. E., Seecretan, L., and Giuliani, J., "Experiments for Satellite Environmental Measurements." In: *Annals of the International Geophysical Year*. London, Pergammon Press, 1958, Vol. 6, pp. 319–321.
2. "Engineering Notes for PM Satellite Transmitter." In: *Check-out Procedures for Missile 16-S Payload.* Redstone Arsenal, February 11, 1959.
3. Heller, G., and Jones, B. P. "Thermal Properties of Explorer VII." In: *Juno II Summary Project Report, Vol. I: Explorer VII Satellite*. NASA Technical Note D-608, July 1961, p. 158.
4. Friichtenich, J. F., "Study of Crater Physics." Final Rept. on Contract No. NAS5-763, Space Technology Labs. Rept. 8980-0003-RU-000, July 18, 1961.
5. Alexander, E. M., McCracken, C. W., et al. "Review of Direct Measurements of Interplanetary Dust from Satellites and Probes." In: *Space Research III: Proc. 3rd Internat. Space Sci. Sympos., Washington, May 1962*. Amsterdam, North-Holland Publ. Co. (To be published).

DUST BOMBARDMENT ON THE LUNAR SURFACE

CURTIS W. McCRACKEN

Goddard Space Flight Center

AND

MAURICE DUBIN

NASA Headquarters

Various types of observational data which are available for the vicinity of the earth and for interplanetary space are reviewed and evaluated in an attempt to establish a good estimate of the flux of small interplanetary dust particles impacting on the moon. A porous, low-density surface layer consistent with photometric and radiometric observations is assumed to exist on the moon, and the effects of the impacting dust particles are considered.

The interplanetary particulate material accreted by the moon amounts to approximately 1 gm/cm^2 for dust particles with masses less than 10^4 gm if the flux has remained constant during the past 4.5×10^9 yr. This value for the accretion rate represents a lower limit if the flux has decreased appreciably since the time of formation of the major lunar surface features.

The porous surface layer acts as a protective covering against hypervelocity impacts of small dust particles and inhibits the production of high speed spray particles which could escape from the moon. The surface layer therefore consists of a mixture of lunar and interplanetary material. The hypervelocity impacts of dust particles constitute an effective mechanism for development and maintenance of a dendroid surface layer of high porosity and low density.

INTRODUCTION AND STATEMENT OF THE PROBLEM

The moon, like the earth, is subjected to a continuing bombardment by interplanetary dust particles having a wide range of sizes. The dust particles impacting on the moon produce markedly different effects, however, than those impacting on the earth. This difference arises because any atmosphere which may be present on the moon is not of sufficiently high density to destroy the particles by vaporization, ablation, or fragmentation (or to decelerate the particles) before they strike the lunar surface. The dust particles impact on the moon at speeds undiminished from their original speeds (relative to the moon). In fact, the speed of a particle at impact is slightly higher than the original speed, as a result of the gravitational acceleration produced by both the moon and the earth.

Dust particles having different origins and different orbital distributions are involved; therefore, the average speed at impact depends (to some extent) on the size of the particle. Typical average speeds are 10 km/sec for dust particles with dimensions of microns, 30 km/sec for particles with dimensions of millimeters, and 15 km/sec for bodies with dimensions of meters. These high speeds lead to hypervelocity impacts, events of explosive violence that result in destruction of both the particle and a portion of the target. The impacting particle and the target material undergo vaporization and fragmentation (or spallation). Most of the fragments and vapors are expelled from the site of the impact. The speeds and ejection angles depend (to a large extent) on the physical structure of the target material and possibly on the structure of the impacting dust particle.

Some of the vaporized and fragmented material may be ejected at speeds greater than the lunar escape speed and, hence, will escape from the moon. The question arises as to whether the

moon is gaining or losing mass as a result of bombardment by dust particles. Negative accretion (net loss of mass) requires that the outer layer of the moon is continually being eroded and removed, thereby exposing new surface material. Positive accretion (net gain of mass) could occur through accretion of interplanetary material but does not necessarily require that no material escapes from the moon. An important consequence of positive accretion would be the formation of a layer of dust and rubble on the lunar surface. If this layer were dendroidal, it could (for certain sizes of impacting dust particles) effectively inhibit the ejection from the moon of fragments and vapors created in a hypervelocity impact.

There are several external mechanisms besides bombardment by interplanetary dust particles which possibly affect the lunar surface in various manners. These mechanisms include bombardment by solar ultraviolet and x-ray radiation, by solar particles (principally protons and electrons), and by cosmic rays. The effects possibly produced by these mechanisms, as well as the nature of the lunar surface, have been discussed by various investigators, including Öpik, Whipple, and Gold.[1–3]

The relative importance of dust bombardment in affecting the microstructure of the lunar surface depends on several aspects of the problem. The fluxes of dust particles of various sizes, the physical structure of the dust particles, the nature of the lunar surface, and the type of hypervelocity impact which applies are important in an investigation of the effects of dust bombardment. The topics of fluxes of dust particles, probable nature of the lunar surface, and possible effects of dust bombardment on the lunar surface are reviewed in subsequent sections, by using presently available data.

FLUXES OF INTERPLANETARY DUST PARTICLES ON THE MOON

Widely different estimates of the accretion rate of interplanetary dust particles have appeared in the literature. It is well to investigate in detail the observational data on fluxes of dust particles, because of the possible importance of dust in determining the lunar surface characteristics.

Information concerning the accretion rates of interplanetary particulate aggregates for the moon is directly available only for those bodies that are sufficiently large to produce visible craters on the moon. The limit on telescopic resolution corresponds to a crater diameter of approximately 750 m (0.4 mi.) for photographic observations and 200 m (0.1 mi.) for visual observations, according to Kuiper.[4] Formation of such a crater requires a particulate aggregate having a radius and a mass of about 10 m and 10^{10} gm, respectively, based on the computations of Öpik.[5]

Statistical studies of the frequencies and size distributions of craters on the moon have been made by several investigators, notably Öpik; Shoemaker, Hackman, and Eggleton; McGillem and Miller;* Palm and Strom; and Kreiter.[6–10] This paper is concerned primarily with the fluxes and effects of particles having masses considerably smaller than those responsible for visible lunar craters, so the subject of crater statistics will not be pursued.

Attention is directed, in general, to bodies having masses less than 10^{10} gm and, in particular, to particles having masses less than 10^{4} gm. Indirect approaches must be employed in establishing the fluxes for particles with masses less than about 10^{10} gm. Data which can be used in these indirect approaches are available for two regions of space—the vicinity of the earth and interplanetary space. These regions are sufficiently far from the moon to make the necessary interpolation subject to some uncertainty. Interpolation toward the moon from the two regions for which data are available does, however, permit fairly realistic limits to be set on the fluxes of interplanetary dust particles impacting on the moon.

The approach in establishing the fluxes of dust particles for the moon follows. Relevant data on the fluxes of dust particles in the vicinity of the earth are compiled to give a cumulative mass distribution that is valid for the vicinity of the earth. Portions of this mass distribution are then considered to be valid also for the vicinity of the moon, and the results from studies of the zodiacal light are used to replace the portion of the dis-

* See also References 11 and 12.

tribution which probably is valid only near the earth.

It is felt that this procedure leads to a cumulative mass distribution which applies in the vicinity of the moon and is valid for dust particles ranging in size from the smallest particles permitted by solar radiation pressure to remain in closed heliocentric orbits up to particles large enough to produce telescopically visible craters on the moon. Of course, the accuracy of this cumulative mass distribution is order-of-magnitude, because of the uncertainties in the observational data used as a basis for the distribution.

The data from which the fluxes are derived are considerably more accurate and extend over a much larger range of particle mass for the vicinity of the earth than for interplanetary space. An attempt has been made to select some of the most reliable data for use in establishing the fluxes for dust particles of different sizes. The choice of data to be included in the compilation has been influenced by the desire to cover as completely as possible the range of particle mass extending from 10^{10} gm down to 10^{-14} gm. It is believed that the selected data is quite representative of most of the data available in the literature.

Data on the fluxes of dust particles in the vicinity of the earth are obtained with a number of observational techniques. The most suitable data come from studies of the frequency of meteorite falls, observations of meteors with photographic, visual, and radar techniques, and direct measurements made with rockets and satellites. These data are the most reliable for use in establishing a mass distribution that can be considered valid for the vicinity of the earth.

Information about the spatial density of dust particles in interplanetary space has been obtained from photometric studies of the zodiacal light. Reasonable assumptions regarding the mass density and average speed for these particles lead directly to estimates of the flux of small interplanetary dust particles impacting on the moon. Also available is a direct measurement of the flux of dust particles in interplanetary space obtained recently by Alexander[13] with the probe Mariner II (1962 $\alpha\rho 1$).

A cumulative mass distribution for the vicinity of the earth (derived on the basis of the selected data) is shown in figure 1. The data used and the assumptions made in constructing the mass distribution are discussed later. In figure 1 the cumulative flux of particles with mass m and larger is plotted as a function of particle mass. The data obtained from the direct measurements and the studies of meteorite falls are plotted directly in terms of particle mass. The data from meteor observations are plotted in terms of visual magnitude, which is related to the mass of a particle in a manner to be described later. A scale of radius of the particles (computed on the basis of a mass density of 1 gm/cm^3) is included for convenience in referring to the approximate size of a particle.

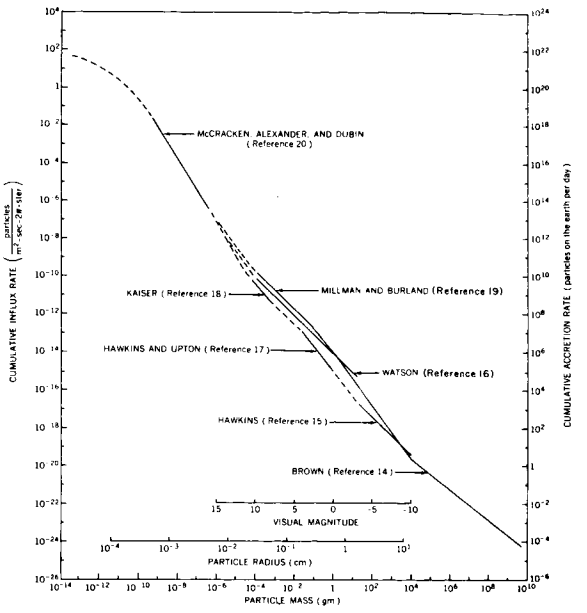


FIGURE 1.—Cumulative mass distribution for interplanetary dust particles in the vicinity of the earth, derived from studies of the frequency of meteorite falls, from observations of meteors, and from direct measurements obtained with rockets and satellites.

The results obtained with different observational techniques are generally expressed in terms of different parameters of the dust particles or in terms of different phenomena produced when the dust particles encounter the atmosphere of the earth. Only the studies of meteorites (and micrometeorites, which are not considered here) directly yield information about the masses of the particles. This information is subject to some uncertainty because of the difficulty in estimating

the amount of material lost through ablation and vaporization as the particles enter the atmosphere.

Brown[21] and Hawkins[22] recently analyzed the frequency of meteorite falls as a function of both the mass of the meteorite and the type of meteorite. The results obtained in these two investigations are in good agreement regarding the rate of infall of large meteorites but are in disagreement about both the fluxes of the small meteorites and the mass distributions of iron and stone meteorites. Only the results of Brown will be considered. Brown gave cumulative mass distributions (in tabular form) for iron and stone meteorites. If the results are combined and the fluxes revised upward by the factor of 3.4 later suggested by Brown,[14] a cumulative mass distribution can be derived for all meteorites incident on the earth. The range of validity is considered here to be 10^4 and 10^{11} gm, although Brown considered the range to be 1 to 10^{11} gm.

The results from direct measurements of fluxes of dust particles with rockets and satellites probably constitute the most reliable data about the fluxes of small dust particles in the vicinity of the earth. These measurements are subject to only small uncertainties (≤ 2) arising from the fact that a value for the average speed of the particles presently must be assumed. Most of the detectors flown have been sensitive to the momentum of an impacting dust particle; hence, an average speed (relative to the detector) must be assumed in order to express the results in terms of particle mass. The available direct measurements have been collected, reviewed, and evaluated in several recent papers.[20,23] The results most probably are not applicable to the moon because of the existence of an enhancement of the fluxes of small dust particles in the vicinity of the earth.[24]

The literature contains a lot of data on the fluxes of dust particles of meteoroidal size. The results come from radar, visual, and photographic observations of both shower and sporadic meteors. Only the data on sporadic meteors need be considered in regard to average fluxes over time intervals longer than 1 yr.

The mass of a particle is not a directly observable parameter in meteor observations; it must be computed on the basis of meteor theory. The observable parameters include electron line density, luminous intensity, and photographic intensity for radar, visual, and photographic meteors, respectively. The observed fluxes for both radar and photographic meteors (for convenience in comparing the results with the visual observations) usually are expressed also in terms of visual magnitude, which is a logarithmic measure of the luminous intensity. These conversions involve some uncertainties, but the major uncertainty is encountered upon trying to express the results from meteor observations in terms of the mass of a particle. The mass-to-magnitude relation is subject to revision as better values for the parameters used in meteor theory become available.

Watson summarizes the results of early visual and telescopic observations of meteors with visual magnitudes extending from -3 to $+10$.[16] The results are given as an incremental magnitude distribution (presented in tabular form) from which a cumulative magnitude distribution can be readily derived.

Millman and Burland reported the results of radar, visual, and photographic meteor observations in Canada.[19] These results indicate that as the brightness increases the number of meteors progressively becomes smaller than would be expected on the basis of the results given by Watson. McKinley has reviewed the Canadian results and given cumulative magnitude distributions (in equation form) derived from the observations.[25] The distributions were extrapolated slightly to cover the range of visual magnitude extending from -10 to $+10$.

Hawkins and Upton[17] analyzed a sample of the observations made in the Harvard Photographic Meteor Program. The results were expressed as cumulative magnitude distributions for both photographic and visual magnitudes. The distribution expressed in terms of visual magnitude will be used here, and the range of validity (as read from the plotted observational data) will be considered to extend from 0 to $+4.1$ in visual magnitude. The cumulative mass distribution given by Hawkins and Upton cannot be used here, because it is based on a different mass-to-magnitude relation than the one adopted for use in this paper.

Hawkins analyzed the available data on fireballs

and bright meteors in an attempt to establish the influx rates[15] and concluded that the number of asteroidal and cometary meteoroids were equal somewhere within the range 0 to −5 in visual magnitude. Hawkins adopted a visual magnitude of −3 as a fiducial point and then used a constant-mass - per - magnitude extrapolation toward brighter meteors (extending to −10 in visual magnitude) as a cumulative magnitude distribution for the fireballs. This same procedure has been followed here by using the cumulative magnitude distribution (expressed in terms of visual magnitude) that was given by Hawkins and Upton.[17] Hawkins also converted the cumulative magnitude distribution to a cumulative mass distribution, but this conversion was based on a mass-to-magnitude relation different from the one adopted here.

Kaiser gave a cumulative flux for radar meteors with radar magnitudes of +10.8 and lower.[18] The gradient of the magnitude distribution was given for meteors with radar magnitudes between +8 and +10.8; a cumulative magnitude distribution can be derived.

The foregoing results from the meteorite studies by Brown, from the meteor observations by Watson, Millman and Burland, Kaiser, Hawkins and Upton, and Hawkins, and from the direct measurements with rockets and satellites were used in constructing the cumulative mass distribution shown in figure 1. All the results from meteor observations were plotted in terms of visual magnitude. (It was assumed that radar and visual magnitude are approximately equal.) Visual magnitude was then related to the particle mass by assuming that a meteoroid with a mass of 1 gm and a speed of 30 km/sec will, on the average, produce a meteor with a visual magnitude of zero. This relation between visual magnitude and particle mass was used recently by Whipple in a revision of the mass-to-magnitude relationship.[26] It is particularly convenient, because it matches the zero point of the visual magnitude with the zero point of the logarithmic mass scale.

The cumulative mass distribution shown in figure 1 is specifically valid for the vicinity of the earth. The segments of the curve which were derived from the studies of meteors and meteorites can be applied readily to the vicinity of the moon, as can be demonstrated by the following argument. The average speed observed for meteors is approximately 30 km/sec. The average speed for the meteorite-producing meteoroids is about 17 km/sec, according to Whipple and Hughes.[27] Such speeds could not have been changed much from the original values by the gravitational effects of the earth. The fluxes of meteoroid-size particles are, therefore, essentially the same for the moon as for the earth. Öpik suggests that the flux at the moon is about 80 percent of that for the earth.[6] Brown comes to a similar conclusion by considering the flux near the earth to be 1.3 times that for the moon.[21] The value given by Öpik is adopted for use here.

Determination of the fluxes of small dust particles presents a problem. As was pointed out, the fluxes measured near the earth with satellites and rockets probably do not apply for the vicinity of the moon. Since no reliable direct measurements have been obtained for the vicinity of the moon, other data must be used for establishing the impact rates of small dust particles on the moon. Data for this purpose are available from studies of the zodiacal light by Allen, van de Hulst, Elsässer, and Ingham.[28–31] Also available is a direct measurement made with the spacecraft Mariner II as reported by Alexander.[13] The results of Elsässer and Ingham are used here in computing the flux of small dust particles on the moon. The direct measurement is in fair agreement with either of these sets of data.

The results from the photometric studies of the zodiacal light were given as incremental size distributions valid within certain limits on the size of the dust particle. These distributions are integrated between the appropriate limits to yield a cumulative size distribution. A mass density of 1 gm/cm^3 and an average speed of 10 km/sec in interplanetary space are applied for conversion from spatial density of particles of a given size and larger to a flux of particles of a given mass and larger which can be used for the vicinity of the moon.

Portions of the cumulative mass distribution shown in figure 1 are applied to the vicinity of the moon and merged with the distributions derived from the results of zodiacal light studies to yield the distribution shown in figures 2 and 3. (The

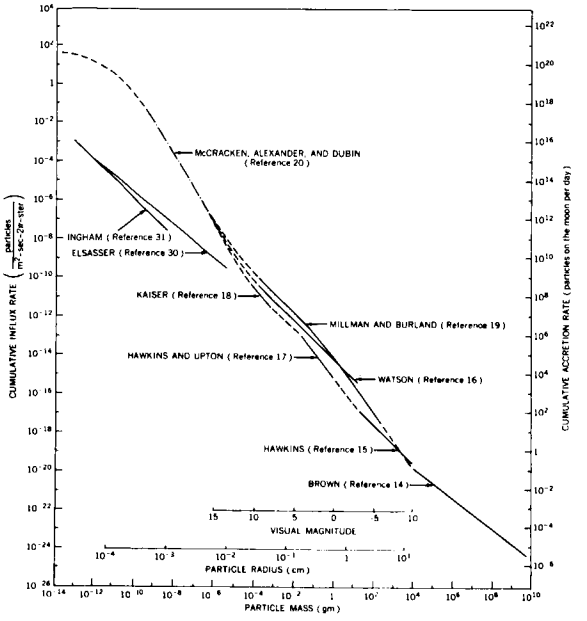


FIGURE 2.—Cumulative mass distribution for interplanetary dust particles in the vicinity of the moon, derived from studies of meteorites and meteors and from photometric studies of the zodiacal light and the solar F corona. (The results obtained with rockets and satellites near the earth are shown for the purpose of comparison.)

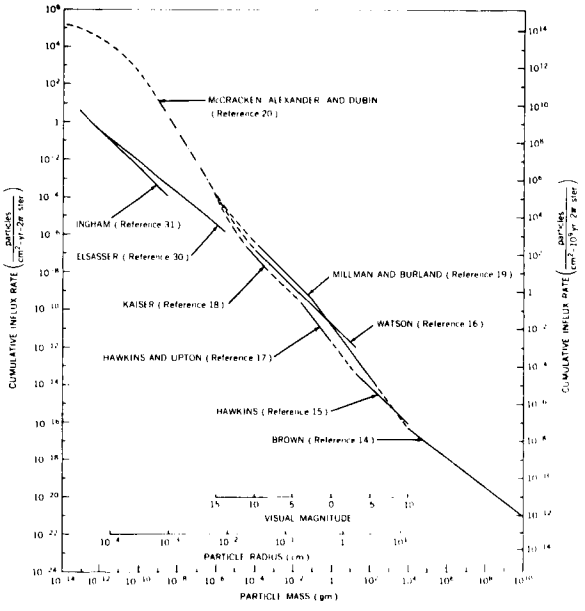


FIGURE 3.—Cumulative mass distribution for the vicinity of the moon. (This figure is identical to figure 2 except for the units on the ordinates.)

flux is expressed in various units in figures 1, 2, and 3 for convenience in making calculations for different intervals of time.) The distribution derived from measurements made near the earth with satellites is shown as a dot-dash curve in figures 2 and 3 for the purpose of comparison. The geocentric distance to which the high fluxes measured near the earth apply is not known; the fluxes of small dust particles on the moon probably fall between the values indicated by the zodiacal light studies and the values indicated by the direct measurements obtained in the vicinity of the earth. The fluxes are, however, thought to be close to those indicated by the zodiacal light studies.

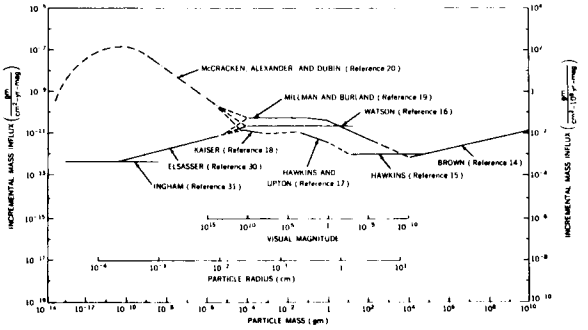


FIGURE 4.—Incremental mass influx distribution for interplanetary dust particles impacting on the moon, derived from the cumulative mass distribution shown in figure 2.

A useful distribution curve (which seems to show some significant trends) can be derived from the cumulative distribution shown in figure 3. Since the derivation involves differentiation, the uncertainties which can occur when a not-too-well-known function is differentiated must be remembered. The cumulative mass distribution is first differentiated with respect to the particle mass to give an incremental mass distribution. The flux of particles having masses within an incremental range of particle mass is then multiplied by the particle mass. This operation results in the incremental mass influx distribution shown in figure 4. The unit range of particle mass is taken to be unit magnitude, because of the extensive use of magnitude (visual) in the meteor literature. An increase of 1 magnitude corresponds to a decrease

of 0.4 in the logarithm of the particle mass and to a decrease by a factor of $100^{1/5}=2.5$ in the mass of the particle. Magnitude has no physical meaning outside the meteoroidal range of particle mass, but it can be used as a convenient logarithmic measure of the mass.

An incremental mass influx curve (similar to the one shown in figure 4) for the vicinity of the earth shows that the process of accretion of material by the earth is dominated by particles of micron size. The incremental mass influx curve for the moon (figure 4) shows that the particles of meteoroid size most probably represent the major contribution in the accretion of interplanetary material by the moon. This conclusion would be affected if the mass of a particle which produces a meteor of zero visual magnitude were reduced appreciably. According to Levin,[32] the value of 1 gm used by Whipple[26] for the mass of a zero visual magnitued meteor should be reduced by 10 to 100 times.

The incremental mass influx curve shown in figure 4 also can be transformed into an energy flux or a momentum flux distribution if an average speed is assumed to hold over most of the range of particle size shown. A value of about 25 km/sec would be appropriate for particles having masses less than about 10^4 gm.

The amounts of interplanetary dust accreted by the moon during a given interval of time can be computed on the basis of the mass distributions shown in figures 2, 3, and 4. The accretion rate can be read directly from figure 4 with an accuracy that is well within the limits of uncertainty in the observational data from which the distributions were derived. (It should be remembered, for convenience in using figure 4, that an increase of 5 in the magnitude corresponds to a decrease by a factor of 100 in the mass of the particle.) Use of the distributions derived from the results given by Hawkins, Hawkins and Upton, Kaiser, Elsässer, and Ingham leads to an accretion rate of approximately 0.5 gm/cm^2 for particles with masses less than 10^4 gm incident on the moon during the past 4.5×10^9 yr. Use of the distributions of Millman and Burland, Elsässer, and Ingham leads to an accretion rate of approximately 3 gm/cm^2 for the same range of particle mass and the same interval of time.

An accretion rate of 1 gm/cm^2 for particles with masses less than 10^4 gm impacting on the moon during the past 4.5×10^9 yr is adopted here as consistent with the available data. According to Öpik[6] the lifetime of interplanetary bodies near the orbit of the earth is about 2×10^6 yr, which means that the flux has probably remained fairly constant during most of the past 4.5×10^9 yr. If the flux were lower at the present time than during the past, the accretion rate would need to be revised upward by an unknown amount.

According to the distribution shown in figure 4, the major contribution to the accreted material comes from particles which would produce radar, visual, and photographic meteors if they were to encounter the atmosphere of the earth. Such particles have masses between about 10^{-5} and 1 gm. Their dimensions are measured in millimeters and tenths of millimeters. Studies conducted at Harvard have shown that these particles are primarily of cometary origin and are very fragile. The mass densities of the particles are of the order of 0.5 gm/cm^3.

A comparison may now be made between the accretion rate derived here (for particles with masses less than about 10^4 gm) and the accretion rates previously suggested by various investigators. Gold and Warner used an accretion rate of 10^{-7} cm^3/cm^2/yr in considering the role of large scale dust erosion on the moon.[33,34] Whipple suggested that the accretion rate was approximately 1 gm/cm^2/2×10^6 yr.[2] Öpik used an accretion rate of 10^{-2} gm/cm^2/10^6 yr in a discussion of the nature of the lunar surface.[6] A significantly lower rate of 10^{-9} gm/cm^2/yr was used by Sharanov.[35] For purposes of comparison with the foregoing estimates note that the accretion rate of 1 gm/cm^2/4.5×10^9 yr may be expressed also as 2×10^{-10} gm/cm^2/yr, 2×10^{-4} gm/cm^2/10^6 yr, 4×10^{-4} gm/cm^2/2×10^6 yr, and (for a mass density of 0.5 gm/cm^3) 4×10^{-10} cm^3/cm^2/yr. The value used here is significantly lower than the values given by Gold, Whipple, and Öpik, but is comparable with the value used by Sharanov.

THE NATURE OF THE LUNAR SURFACE

An investigation of the effects of dust bombardment on the moon requires that adequate consideration be given to the physical characteristics

of the lunar surface, especially on a small scale. Some parameters which are important in determining the effects of bombardment by dust particles are the composition and physical nature of the outermost lunar surface layer, the size of the surface features relative to the size of an impacting dust particle, and the interstitial nature of the grain distribution (or the porosity of the surface layer). These parameters specify the bulk density and other characteristics of the surface material. The density and structure of the impacting dust particles are also relevant in the interaction.

Some physical properties of the lunar surface have been determined by using information obtained from radar, visual, infrared, microwave, photometric, and polarimetric observations of the moon. The known properties (such as the albedo, the heat conductivity, the dielectric constant, the degree of roughness, the color, and the photometric and polarimetric properties) impose restrictions on the physical nature of the lunar surface layer. The surface can be neither smooth rock nor compacted rubble but rather must be composed of small particles arranged to have properties consistent with the observational information.

The lunar surface characteristics may be simulated by a dendriform material of low bulk density, high porosity, and a depth ranging from centimeters to several meters. The surface structure has probably developed because of the adhesive properties of micron-diameter dust particles and gas condensation in an ultrahigh vacuum. Solar radiation in the extreme ultraviolet and soft x-ray region and proton bombardment may have contributed to the adhesive properties.

The results of many of the studies of lunar surface characteristics have been reviewed and evaluated by Fessenkov, Dollfus, Sinton, and Evans.[36–39] Öpik, Whipple, Gold, Sharanov, Sytinskaya, Troitskii, Salisbury, and others have considered the surface structure of the moon.[1–3,33,35,40–42]

Radar observations of the moon at metric and decimetric wavelengths give information about the structure of the surface on a smaller scale than is possible with telescopic observations. Evans has recently reviewed the lunar radar results.[39] Important conclusions based on radar studies are that the lunar surface is smooth and undulating with gradients of less than 1 in 10 at meter wavelengths and that less than 10 percent of the surface is covered by outcroppings with dimensions of the order of a meter. The surface is smooth down to dimensions of approximately 10 cm. Also important are the low values for the dielectric constant obtained from radar studies. These values (ranging between 1.1 and 2.7) are lower than those for ordinary terrestrial rocks, indicating a surface material of low bulk density and of a depth between a few cm and 10 m.

Results from photometric and polarimetric studies of the moon and terrestrial materials have been discussed by Minnaert, Fessenkov, and Dollfus.[36,37,43,44] Some general conclusions about the lunar surface model reached on the basis of the photometric and polarimetric studies, which are important in considerations of the effects of dust bombardment, are as follows: The photometric function is nearly the same for all regions of the lunar surface, even over the range in albedo existing between the highlands and maria. Color differences on the moon are small. The moon shows no limb darkening—all points on the lunar surface reach maximum radiance near the time of a full moon. The polarization is nearly constant in magnitude and direction for the highlands but varies irregularly over the maria. The polarization varies approximately inversely with albedo. The photometric properties suggest a porous surface layer, existing even on the steepest slopes. The polarimetric results are interpreted by Dollfus and Minnaert to imply that the porous layer is covered with dust. Fessenkov is more explicit, suggesting that the surface layer may consist of agglomerations of grains.

Sinton and Zel'tser have reviewed the results of infrared measurement of the lunar surface temperature.[38,45] Analyses of the temperature variation of the lunar surface lead to the conclusion that most of the surface is covered with a dust layer with a thickness greater than 5 cm.

Attempts have been made to find ordinary terrestrial materials having photometric and polarimetric properties similar to those of the moon. The attempts met with little or no success until recently. According to Minnaert,[43] van

Diggelen found that the photometric function of a lichen closely resembled the photometric function of the moon. An examination of the fairy castle or dendriform nature of the lichen gives an impression of how really complex the microstructure on the surface of the moon may possibly be.

The probable complexity of the lunar surface has been demonstrated more vividly in the recent successful attempts of Hapke to construct surfaces by sifting various powdered minerals.[46] Using common terrestrial materials Hapke created surfaces which show photometric properties similar to those of the lunar surface. The mineral grains tend to build highly complex structures of dendroidal form. The surfaces have irregular interstices. These experiments are the first to succeed in giving an arrangement of materials which have photometric properties similar to those of the lunar surface and which are likely to be present on the moon. The structures may also have thermal conduction properties which would explain the lag in cooling during the lunar night. Radiation damage to the mineral crystals by solar protons has been suggested as an agent for discoloring the crystals to the required opacity.

In summary, the results from the microwave, infrared, and photometric studies of the moon indicate that the surface layer is everywhere a low density porous material. The polarimetric studies indicate that a thin layer of dust exists, or (more probably) that small structures or agglomerations of particles (dendroids) cover the lunar surface. A surface layer of the complex dendriform structures created in the laboratory by Hapke will therefore be adopted in this paper as a working model for use in investigating the effects of dust bombardment on the lunar surface. It is assumed that the grain size is microscopic (of the order of 10 μ) and that the grains are agglomerated into a dendroid structure. The dendroid nature of the agglomerates of dust particles results in a material of very low bulk density, probably 10 times less than the density of the parent material. The structure is effectively very open with interstitial spaces of various sizes and shapes. A depth of a few cm for the layer will serve as a starting point in the investigation.

Yet remaining, however, is the hypothesis of a mechanism on the lunar surface to act as the analog of the sifting process used in the laboratory to distribute the grains. The process must be capable of developing and maintaining the complex structures at a rate which would keep the lunar surface in a state of good repair from a photometric standpoint. Yet the mechanism must not be so dominant that it erases all photometric and polarimetric differences on the moon during a time scale of several billion years.

HYPERVELOCITY IMPACTS ON THE LUNAR SURFACE

The dendroid model of the lunar surface (described in the previous section) may have evolved during the early history of the moon at the time of formation of the maria and large craters. Following this period, dust particle bombardment could have affected the structure of the lunar surface. This is a process similar to that required to build the lunar surface. Very likely a quasi-equilibrium microstructure on the lunar surface has developed and been maintained up to the present time, if a single generation process is responsible. Although mechanisms such as volcanism and lava flows have been considered as the source of a number of lunar surface characteristics, impacts from large bodies are clearly the source of most of the craters on the moon. On a smaller scale, hypervelocity impacts by dust particles have had an effect on the surface characteristics of the moon.

The speeds (from 2.4 to 70 km/sec) for impacts of extralunar particles are catastrophically high—impacts at these velocities are called hypervelocity impacts. Speeds at the lower end of this range may be simulated experimentally by using special light-gas guns and chemical and electrical means for particle acceleration. The upper limit of speeds which have been obtained experimentally in studies of cratering is about 8 to 10 km/sec, although higher speeds can be obtained in some circumstances.

Most of the experimental data on hypervelocity impacts apply for metal targets and dense projectiles. These data are inadequate for determining the effects of impacts on the moon because great differences may be expected between hypervelocity cratering on a metal target and on a surface with the characteristics of the lunar

surface. The hypervelocity interaction also depends on the density and properties of the impacting projectile. Experimental simulation of projectiles corresponding to interplanetary dust particles has not been achieved. The model for these particles is not adequately known, and the required fragility of the projectile cannot be simulated and maintained during acceleration. Some experimental data for hypervelocity impacts on rock surfaces have been obtained. The results from cratering on basalt targets are different from cratering effects observed for metals. The effects of hypervelocity impacts on sand or dust targets (and particularly on lunar-like surfaces) might be expected to differ markedly from the effects of impact on metal or basalt. The possible differences are discussed below.

A hypervelocity impact on a surface of metal or basalt forms a crater and generates high speed ejecta during the cratering process. The crater formed in a metal target is generally hemispherical; it has an edge or lip extending above the level of the target surface, and the diameter of the crater is several times greater than the diameter of the projectile. A crater is formed in a basalt target in the same manner as in a metal target, but the effect of spalling or fracturing (resulting from shock propagation in the target material) is much more pronounced. Spallation in basalt greatly increases the amount of ejecta (compared with metal targets) and accordingly increases the volume of the crater. The volume of the crater formed in sand has been found by Gault, Shoemaker, and Moore to be still larger than that formed in basalt.[47]

Öpik, Bjork, and Stanyukovich and Bronshten have made theoretical studies of hypervelocity impacts in metals and rocks.[5,48,49] Bjork has probably gone furthest in solving the theory of hypervelocity cratering in metals by using a hydrodynamic model. The mechanism of crater formation is essentially one of cavitation, resulting from an intense plastic deformation wave formed during the impact. The size and shape of the crater are determined by the properties of the wave and the target material. The initial conditions applying during the early stages of the impact determine the amplitude and shape of the deformation wave. The initial stage of the impact or primary penetration is characterized by a very rapid plastic deformation of the target and impacting projectile. If the speed at impact is very high, the surface pressure so far exceeds the yield strength of the material that a hydrodynamic treatment is quite accurate. The deformation wave propagates into the target, displacing material as it disperses. The effects produced in the target depend upon the properties of the target material as described by the equation of state of the material. The crater dimensions are determined by the distance traveled by the deformation wave while its intensity is greater than the strength of the target material. The model described is essentially the same as that used by Bjork in his theoretical treatment of cratering.

The hydrodynamic model for hypervelocity impacts represents a fairly accurate approximation for impacts on metal targets. Experimental data show agreement with this model for speeds up to 11 km/sec. For the case for impacts on inelastic or brittle surface materials, the effect of shock propagation through the material results in a degree of spallation that depends on the frangibility of the target material. Experimental data for hypervelocity impacts on basalt have been obtained by Gault, Shoemaker, and Moore.[47] Craters formed by hypervelocity impacts of projectiles at velocities of 6.4 km/sec on basalt were studied in detail. The volume of the crater, expressed in terms of the cumulative mass of the material ejected from the surface, was about 200 times the projectile mass. The ejected mass was more than 10 times greater then that observed for high speed impacts in metals and represents the effects of spallation in a brittle material. The size and velocity distributions of the ejected particles were measured.

In addition to impacts on basalt, hypervelocity impacts were made on weakly bonded quartz sand. It was found that for a given expended energy, the mass ejected from the bonded sand target was at least a factor of 3 greater than the mass ejected from the basalt target. The velocity distributions of the ejected particles did not differ very much from those for basalt. For impacts on sand targets the ejected particles included grains of sand, agglomerations of sand grains, and finely crushed quartz.

Reasons exist for believing that the effects of hypervelocity impacts on a surface of dendroidal structure would be considerably different from those of impacts on metals, rocks, sand, and dust. The major difference for the dendroid model arises from the fact that this model has a compressibility. The bulk density of the dendroidal structures created by Hapke is a factor of 10 less than the actual density of the grains. An analogy for impacts into a compressible structure would be the impacts of meteoroids into a gas such as the atmosphere of the earth. In this extreme case no ejecta result, because all of the momentum is absorbed in this essentially inelastic medium. Similar effects have been observed experimentally for hypervelocity impacts into materials such as styrofoam. Projectiles with speeds of a few km per sec can be completely stopped and recovered intact. At higher speeds the projectile may fracture, but the pieces are trapped in the compressible porous matrix. Very little material is ejected by such impacts. Hypervelocity impacts and the resulting craters in metal, basalt, and a dendroidal structure are depicted in figure 5. In the first two cases a considerable volume is ejected from the surface and the crater retains the shape formed at impact. For impacts into a dendroidal structure the projectile can penetrate deep below the surface and, while fracturing, can disperse laterally into the medium. The energy and momentum absorbed tend to expand or blow out the medium. It appears reasonable to believe that the impact will result in the raising of the surface in the vicinity of the impact without necessarily forming a crater. The hypervelocity impact would thereby lead to the generation of more small particles for building the dendroidal structure.

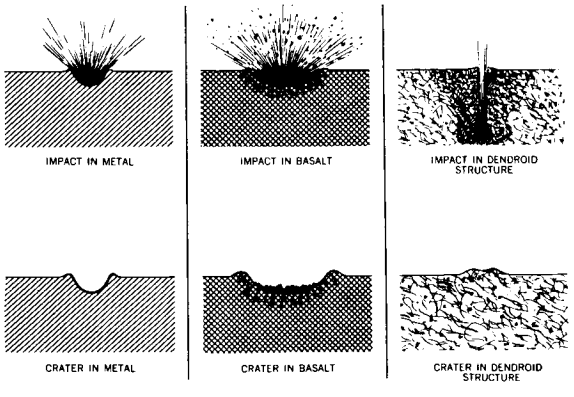


FIGURE 5.—Hypervelocity impacts on various surfaces.

CONCLUSIONS

The nature of the lunar surface layer as related to the impacts of interplanetary dust particles has been an enigma. The dendroidal surface model created by Hapke successfully reproduces many of the properties of the lunar surface. It incorporates the adhesive or cemented qualities of the weak, porous matrix discussed by Whipple. It protects the underlying surface against erosion by dust particle bombardment as was suggested by Öpik for a dust layer which was not removed. Hapke has pointed out that removal of surface material by a hypervelocity impact into a dendroid layer would be inhibited.

The available data on the fluxes of interplanetary dust particles with masses less than about 10^4 gm show that the material accreted by the moon during the past 4.5×10^9 yr amounts to approximately 1 gm/cm^2 if the flux has remained fairly constant. This value for the accretion rate is significantly lower than several previous estimates. The relative importance of atomic sputtering should be further investigated, because of this low value for the accretion rate.

The hypervelocity impacts of small dust particles constitute an effective mechanism for developing and maintaining a dendroid layer that is consistent with observational data for the lunar surface. The dendroid surface is probably rough on a scale that is comparable with the dimensions of the particles which dominate the accretion process. The ejecta from hypervelocity impacts on such a low density porous structure would be largely captured and retained by the surface layer, leading to positive accretion for the moon.

The lunar surface layer thus formed would consist of a mixture of lunar material and interplanetary material (primarily of cometary origin) from 10 cm to 1 m thick. The low value for the accretion rate for the small particles is not adequate to produce large scale dust erosion or to form deep layers of dust on the moon, for the flux has probably remained fairly constant during the past several billion years.

REFERENCES

1. ÖPIK, E. J., "Surface Properties of the Moon," in: *Progress in the Astronautical Sciences*, ed. by S. F. Singer, Amsterdam: North-Holland Publ. Co., 1962, vol. 1, pp. 215–260.
2. WHIPPLE, F. L., "On the Lunar Dust Layer," in: *Vistas in Astronautics*, ed. by M. Alperin and H. F. Gregory, New York: Pergamon Press, 1959, vol. 2, pp. 267–272.
3. GOLD, T., "Dust on the Moon," in: *Vistas in Astronautics*, ed. by M. Alperin and H. F. Gregory, New York: Pergamon Press, 1959, vol. 2, pp. 261–266.
4. KUIPER, G. P., "The Moon," in: *The Exploration of Space*, ed. by R. Jastrow, New York: Macmillan, 1960, pp. 70–76.
5. ÖPIK, E. J., "Meteor Impact on Solid Surface," *Irish Astronom. J.* **5**(1): 14–33, 1958.
6. ÖPIK, E. J., "The Lunar Surface as an Impact Counter," *Mon. Not. Roy. Astronom. Soc.* **120**(5): 404–411, 1960.
7. SHOEMAKER, E. M., HACKMAN, R. J., and EGGLETON, R. E., "Interplanetary Correlation of Geologic Time," in: *Advances in Astronautical Sciences*, ed. by Horace Jacobs, New York: Plenum Press, vol. 8, pp. 70–89.
8. MCGILLEM, C. D., and MILLER, B. P., "Lunar Surface Roughness from Crater Statistics," *J. Geophys. Res.* **67**(12): 4787–4794, November 1962.
9. PALM, A., and STROM, R. G., "The Craters in the Lunar Walled Plain Ptolemaeus," *Planet. Space Sci.* **11**(2): 125–134, February 1963.
10. KREITER, T. J., "A Possible Method of Dating Lunar Surface Features," Jet Propulsion Lab., Calif. Institute Tech. TR 34–57, April 11, 1960.
11. SALISBURY, J. W, "Discussion of a Paper by C. D. McGillem and B. P. Miller, 'Lunar Surface Roughness from Crater Statistics,'" *J. Geophys. Res.* **68**(7): 2065, April 1963.
12. MCGILLEM, C. D., and MILLER, B. P., "Reply" (reply to Reference 11) *J. Geophys. Res.* **68**(7): 2066, April 1963.
13. ALEXANDER, W. M., "The Mission of Mariner II: Preliminary Observations. Cosmic Dust," *Science* **138**(3545): 1098, December 7, 1962.
14. BROWN, H., "Addendum: The Density and Mass Distribution of Meteoritic Bodies in the Neighborhood of the Earth's Orbit," *J. Geophys. Res.* **66**(4): 1316–1317, April 1961.
15. HAWKINS, G. S., "The Relation Between Asteroids, Fireballs and Meteorites," *Astronom. J.* **64**(10): 450–454, December 1959.
16. WATSON, F. G., "Between the Planets," Cambridge: Harvard Univ. Press, 1956.
17. HAWKINS, G. S., and UPTON, E. K. L., "The Influx Rate of Meteors in the Earth's Atmosphere," *Astrophys. J.* **128**(3): 727–735, November 1958.
18. KAISER, T. R., "The Determination of the Incident Flux of Radio-Meteors. II. Sporadic Meteors," *Mon. Not. Roy. Astronom. Soc.* **123**(3): 265–271, 1961.
19. MILLMAN, P. M., and BURLAND, M. S., "Highlights of Some Papers Presented at the 96th Meeting of the American Astronomical Society," *Sky and Telescope* **16**(5): 222, March 1957.
20. MCCRACKEN, C. W., ALEXANDER, W. M., and DUBIN, M., "Direct Measurement of Interplanetary Dust Particles in the Vicinity of the Earth," NASA Technical Note D-1174, July 1962; also *Nature* **192**(4801): 441–442, November 4, 1961.
21. BROWN, H., "The Density and Mass Distribution of Meteoritic Bodies in the Neighborhood of the Earth's Orbit," *J. Geophys. Res.* **65**(6): 1679–1683, June 1960.
22. HAWKINS, G. S., "Asteroidal Fragments," *Astronom. J.* **65**(5): 318–322, June 1960.
23. ALEXANDER, W. M., MCCRACKEN, C. W., SECRETAN, L., and BERG, O. E., "Review of Direct Measurements of Interplanetary Dust from Satellites and Probes," to be published in: *Proc. Third Internat. Space Sciences Symp.*, North Holland Publ. Co.
24. DUBIN, M., and MCCRACKEN, C. W., "Measurements of Distributions of Interplanetary Dust," *Astronom. J.* **67**(5): 248–256, June 1962.
25. MCKINLEY, D. W. R., "Meteor Science and Engineering," New York: McGraw-Hill, 1961.
26. WHIPPLE, F. L., "On Meteoroids and Penetration," paper presented at the Interplanetary Missions Conf., 9th Annual Amer. Astronaut. Soc. Meeting, Los Angeles, Calif., January 15–17, 1963.
27. WHIPPLE, F. L., and HUGHES, R. F., "On the Velocities and Orbits of Meteors, Fireballs, and Meteorites," in: *Meteors*, ed. by T. R. Kaiser, New York: Pergamon Press, 1955, pp. 149–156.
28. ALLEN, C. W., "The Spectrum of the Corona at the Eclipse of 1940 October 1," *Mon. Not. Roy. Astronom. Soc.* **106**(2): 137–150, 1946.
29. VAN DE HULST, H. C., "Zodiacal Light in the Solar Corona," *Astrophys. J.* **105**(3): 471–488, May 1947.
30. ELSÄSSER, H., "Die raumliche Verteilung der Zodiakallichtmaterie," *Zeit. f. Astrophys.* **33**(4): 274–285, 1954.
31. INGHAM, M. F., "Observations of the Zodiacal Light from a Very High Altitude Station. IV. The Nature and Distribution of the Interplanetary Dust," *Monthly Not. Roy. Astronom. Soc.* **122**(2): 157–175, 1961.
32. LEVIN, B., "Fragmentation of Meteor Bodies," *Nature* **196**(4854): 527–528, November 10, 1962.
33. GOLD, T., "The Lunar Surface," *Mon. Not. Roy. Astronom. Soc.* **115**(6): 585–603, 1955.
34. WARNER, B., "The Lunar Maria," *Planet. Space Sci.* **5**(4): 283–285, August 1961.
35. SHARANOV, V. V., "The Nature of the Lunar Surface," in: *The Moon*, ed. by A. V. Markov, Chicago: Univ. of Chicago Press, 1962, pp. 338–372.
36. FESSENKOV, V. G., "Photometry of the Moon," in: *Physics and Astronomy of the Moon*, ed. by Z. Kopal, New York: Academic Press, 1962, pp. 99–130.

37. DOLLFUS, A., "The Polarization of Moonlight," in: *Physics and Astronomy of the Moon*, ed. by Z. Kopak, New York: Academic Press, 1962, pp. 131–159.
38. SINTON, W. M., "Temperatures on the Lunar Surface," in: *Physics and Astronomy of the Moon*, ed. by Z. Kopal, New York: Academic Press, 1962, pp. 407–428.
39. EVANS, J. V., "Radio Echo Studies of the Moon," in: *Physics and Astronomy of the Moon*, ed. by Z. Kopal, New York: Academic Press, 1962, pp. 429–479.
40. SYTINSKAYA, N. N., "New Data on the Meteoric-Slag Theory of the Formation of the Outer Layer of the Lunar Surface," *Soviet Astronom. AJ* **3**(2): 310–314, March–April 1959.
41. TROITSKII, V. S., "Nature and Physical State of the Surface Layer of the Moon," *Soviet Astronom. AJ* **6**(1): 51–54, July–August 1962.
42. SALISBURY, J. W., "Lunar Surface Characteristics," paper presented at the Automotive Engineering Congress, Detroit, Michigan, January 14–18, 1963.
43. MINNAERT, M., "Photometry of the Moon," in: *Planets and Satellites*, ed. by G. P. Kuiper and B. M. Middlehurst, Chicago: Univ. of Chicago Press, 1961, pp. 213–248.
44. DOLLFUS, A., "Polarization Studies of Planets," in: *Planets and Satellites*, ed. by G. P. Kuiper and B. M. Middlehurst, Chicago: Univ. of Chicago Press, **1961**, pp. 343–399.
45. ZEL'TSER, M. S., "The Temperature of the Lunar Surface," in: *The Moon*, ed. by A. V. Markov, Chicago: Univ. of Chicago Press, 1962, pp. 175–203.
46. HAPKE, B. W., "Experiments Relating to the Lunar Surface. I. Photometric Studies. II. Proton Bombardment of Minerals. Preliminary Report No. 2.," Center for Radiophysics and Space Research, Cornell Univ. Ithaca, N. Y., CRSR-127, July 1, 1962.
47. GAULT, D. E., SHOEMAKER, E. M., and MOORE, H. J., "Spray Ejected from the Lunar Surface by Meteoroid Impact," NASA Technical Note D-1767, April 1963.
48. BJORK, R. L., "Effects of a Meteoroid Impact on Steel and Aluminum in Space," in: *Proc. of the 10th Internat. Astronaut. Cong.*, ed. by F. Hecht, Vienna: Springer-Verlag, 1960, vol. 2, pp. 505–514.
49. STANYUKOVICH, K. P., and BRONSHTEN, V. A., "The Role of External Cosmic Factors in the Evolution of the Moon," in: *The Moon*, ed. by A. V. Markov, 304–337, Chicago: Univ. of Chicago Press, 1962, pp. 304–337.

REARRANGEMENT COLLISIONS. III. ELECTRON TRANSFER FROM ATOMS TO FAST PROTONS*

M. H. MITTLEMAN†

Goddard Space Flight Center

The high energy rearrangement collision, in which a fast proton incident upon an atom picks up an electron, is considered. It is found that with the present understanding of rearrangement collisions the cross section can be given for any atom in a restricted energy region with some theoretical uncertainty due to higher order collisions.

INTRODUCTION

The collision in which a fast proton hits a neutral atom in its ground state and picks up an electron to form hydrogen and a residual ion is considered here.

Previous theoretical treatments of this problem have yielded cross sections the energy dependence of which is $E^{-11/4}$ (Thomas 1927) and E^{-3} (Bohr 1940). Both of these workers considered a model in which the collective effects in the atom played an important role. Brinkman and Kramers (1930) have obtained a result for this cross section with a high energy behavior of E^{-6} from a single-particle model. This is essentially the result we obtain; however, we shall show that this dependence obtains in a very restricted high energy region where no experiments are available.

The present state of the theory of rearrangement collisions is so poor that even the proper starting point for the formulation of the theory is in doubt. That is, there are many expressions for the transition amplitude describing the collision, all of which are exact. These all contain some quantity (a wave function or propagator) which presupposes a knowledge of the exact solution of the problem so that approximations are necessary. Practical considerations usually limit the approximations to a Born approximation or a distorted wave Born approximation for the wave functions. When these approximations are applied to the various expressions for the T matrix different results are obtained. It is one of the purposes of this paper to investigate the difference in the various formulations.

In the next section we present the various formal formulations of the T matrix. In Section 3 we shall see that the first Born approximation to the various formulations leads to different results but that the leading term in an expansion in energy^{-1} in each result yields a cross section which varies as E^{-6} with the same coefficient for each one. The coefficient of the next term (E^{-7}) depends upon the starting formulation. In Section 4 we briefly discuss the second Born approximation. The discussion will be based on previous work where it was shown that the second Born approximation contributes a term in the cross section with energy dependence $E^{-11/2}$ but with a small coefficient. Moreover, there are terms in E^{-6} from the second Born approximation and there are even some indications that the third Born approximation will contribute terms of order E^{-6}. Therefore, even the leading term which seems to be predicted unambiguously from the first Born approximation is in doubt. The net result is that we can only make an approximate prediction of the cross section in a very restricted energy region.

* Supported in part by the U. S. Atomic Energy Commission.

† Permanent address: Lawrence Radiation Laboratory, Livermore, California, and University of California, Berkeley, California.

FORMAL STARTING POINTS

There are many expressions for the transition amplitude for the process:

$$H^+ + A(0) \rightarrow H(n) + A^+(m). \tag{1}$$

Here the symbols in parentheses indicate the state of the body. The difficulties all seem to arise from the fact that the non-interacting initial and final states are described by different Hamiltonians, so that these states are not orthogonal. This lack of orthogonality has led to much misunderstanding of the role of the distortion potential in a distorted wave calculation of such a process.

For instance, in the case of electron transfer from hydrogen to protons the role of the proton–proton interaction has been much discussed. In the original high energy first Born approximation calculation (Brinkman and Kramers 1930) the interaction was completely omitted. Subsequent calculations (Jackson and Schiff 1953, Bates and Dalgarno 1952) have included it as part of the interaction in the first Born approximation. A second Born approximation calculation (Drisko 1955†) shows that the second-order contributions of this potential cancel the first-order ones. Indeed, one can show (Mittleman 1961 a) that in the limit $m/M \rightarrow 0$ it makes no contribution at all.

Historically, the first formal difficulty in the theory of rearrangement collisions was the so-called "post prior" discrepancy. This resulted from the two forms obtained by considering the transition to be caused either by the initial or final perturbation (Lippmann 1956):

$$T_1 = \langle \Psi_f^{(-)} V_i \phi_i \rangle \tag{2}$$

$$T_2 = \langle \phi_f V_f \Psi_i^{(+)} \rangle. \tag{3}$$

Our notation is the following: H is the total Hamiltonian. It may be broken up into two forms, appropriate to the initial or final states

$$H = H_i + V_i = H_f + V_f \tag{4}$$

where $H_i(H_f)$ is the initial (final) non-interacting Hamiltonian and $V_i(V_f)$ is the interaction in the initial (final) state. ϕ_i and ϕ_f are the initial and final non-interacting states respectively. They satisfy

$$(E - H_i)\phi_i = 0 \tag{5a}$$

$$(E - H_f)\phi_f = 0 \tag{5b}$$

where E is the total energy. The state $\Psi_i^{(+)}$ is the solution to the full scattering problem with ϕ_i as the initial state:

$$\Psi_i^{(+)} = \phi_i + (a_i)^{-1} V_i \Psi_i^{(+)} \tag{6}$$

where $a_i = E - H_i + i\eta$ and $\Psi_f^{(-)}$ is the solution to the full time reversed scattering problem with ϕ_f as the initial state (Gell-Mann and Goldberger 1953):

$$\Psi_f^{(-)} = \phi_f + (a_f^+)^{-1} V_f \Psi_f^{(-)} \tag{7}$$

where $a_f = E - H_f + i\eta$.

The forms T_1 and T_2 are formally equal but when approximations are made for the total wave functions the results for T_1 and T_2 are no longer equal. If the correct first Born approximations for $\Psi_i^{(+)}$ and $\Psi_f^{(-)}$ are substituted in each then it can be shown that the results are still equal (Schiff 1949). Higher Born approximations lose this property. This is not surprising in that different potentials are assumed to be small in each case. There is no reason to expect an expansion in which the proton–atom interaction (initial) is considered small to be the same as one in which the hydrogen–ion interaction (final) is taken to be small.

T_2 and T_3 can be modified by replacing ϕ_i and ϕ_f by distorted wave states. If we define the distorted wave states $\chi_i^{(+)}$ and $\chi_f^{(-)}$ by

$$(E + i\eta - H_i - v_i)\chi_i^{(+)} = i\eta\phi_i \tag{8a}$$

$$(E - i\eta - H_f - v_f)\chi_f^{(-)} = -i\eta\phi_f \tag{8b}$$

then the transition amplitudes may be written

$$T = \langle \Psi_f^{(-)}(V_i - v_i)\chi_i^{(+)} \rangle = \langle \chi_f^{(-)}(V_f - v_f)\Psi_i^{(+)} \rangle \tag{9}$$

where v_i and v_f are almost arbitrary potentials. If they are chosen to be potentials defined by the property that they give the exact elastic scattering, then (9) can be written (Mittleman 1961b)

$$T_3 = \langle \Psi_f^{(-)} \mid [V_i, \Pi_i] \mid \Psi_i^{(+)} \rangle \tag{10}$$

$$T_4 = \langle \Psi_f^{(-)} \mid [\Pi_f, V_f] \mid \Psi_i^{(+)} \rangle. \tag{11}$$

† Unpublished thesis, Carnegie Institute of Technology.

Here Π_i is a projection operator onto the ground state of the atom, or any collection of bound states containing the ground state. Π_f projects on to the m state of the ion and the n state of hydrogen or any collection of such bound states containing the product (n, m) (Mittleman 1962).

The re-ordering which changes the pair T_1 and T_2 into T_3 and T_4 is accomplished at the expense of losing the equality of the first Born approximations in the post and prior forms. This is not a great loss in that there does not seem to be any physical reason to expect such a symmetry in the first distorted wave Born approximation. On the other hand, T_3 and T_4 have the appealing property of having transitions only between orthogonal states (Mittleman 1961b).

Another form of T has been obtained by Greider (1959). This is accomplished by singling out the interaction between the proton and the electron which is stripped to form hydrogen. The remaining part of V_i is absorbing as a generalized distortion potential. This potential can cause excitations and even ionizations of the atom, but no rearrangement into hydrogen.† If we now define the wave function $\lambda_i^{(+)}$ by

$$[E+i\eta-H_i-(V_i-V_1)]\lambda_i^{(+)}=i\eta\phi_i \qquad (12)$$

where V_1 is the electron–proton interaction mentioned above, then Greider has shown that an equivalent form for T is

$$T_5=\langle\Psi_f^{(-)}V_1\lambda_i^{(+)}\rangle. \qquad (13)$$

We could go one step further by including another arbitrary potential in (12) and choosing it to be the generalized distortion potential for $\lambda_i^{(+)}$ (Mittleman and Pu 1962). We would then obtain the form

$$T_6=\langle\Psi_f^{(-)}\,|\,[V_1,\,\Pi_1]\,|\,\Psi_i^{(+)}\rangle \qquad (14)$$

where now Π_1 is the extremely complicated projection operator which projects onto $\lambda_i^{(+)}$.

We also mention the variational technique as used by Bates (1959) in which the wave function is assumed to be a linear superposition of the initial and final states with unknown coefficients. These are determined from a variational principle. The remarks concerning the results of Section 3 apply to this method also.

Finally, we should point out that there is an extensive literature dealing with this problem (mostly for atomic hydrogen). No attempt has been made here to provide a complete list of such references. The reader is referred to the review article by Bates and McCarroll (1962) for the latest compilation of these. The most recent calculation of this sort (Bransden and Cheshire 1962†) is an impulse approximation calculation. It contains second-order effects and essentially confirms the results of Drisko.

FIRST BORN APPROXIMATIONS

We now turn to the evaluation of these forms in first Born approximations. We mean by this that any wave function such as $\Psi_i^{(+)}$, $\Psi_f^{(-)}$ or $\lambda_i^{(+)}$ is replaced by its first Born approximation. The approach will be to describe the relative motion of the nuclei classically and to make the further approximation that the motion is unaccelerated (Frame 1931, Mittleman 1961a). First we consider the form T_5. The first Born approximation is (the limit where the electron proton mass ratio is zero is taken)

$$T_5=i\int dt\,d^3y_1,\cdots,d^3y_s\phi_n^*(\mathbf{y}_1-\mathbf{R})\,u_m^{(+)*}(\mathbf{y}_2\cdots)$$

$$\times\frac{-2}{|\mathbf{y}_1-\mathbf{R}|}\,u_0(\mathbf{y}_1,\mathbf{y}_2\cdots)\exp(-\tfrac{1}{2}i\mathbf{y}_1\cdot\dot{\mathbf{R}}+i\delta_{mn}t). \qquad (15)$$

Here we have used units in which length is measured in Bohr radii a, energy in units of the ionization energy of hydrogen $|W_0|$ and time in units

† The only requirement on the distortion potentials here and in (8) is that they be unable to support a binding in the rearranged chanel. A sufficient condition is that there be no proton-stripped electron attraction in this potential.

† Preprint.

of $\hbar/|W_0|$. Equation (15) is written in the rest frame of the atom. The wave functions ϕ_n, $u_m^{(+)}$ and u_0 are those of hydrogen in the nth state, the ion in the mth state, and the atom in the ground state respectively. The internuclear coordinate $\mathbf{R}$ is given by $\mathbf{R}=\mathbf{b}+\hat{a}_z Vt$, where $\mathbf{b}$ is the impact parameter, V is the proton laboratory velocity in units of $e^2/2\hbar$ and δ_{mn} is the change of energy between initial and final states:

$$\delta_{mn}=\epsilon_m{}^+ + W_n - \epsilon_0 + \tfrac{1}{4}V^2. \tag{16}$$

We shall make the approximation that the atom and ion are both describable by Hartree–Fock wave functions and that the orbitals are derived from the same Hamiltonian in each case. This approximation is better the heavier the atom. It is poor in the outer regions but we shall see that the main contribution to our result comes from the inner regions. With this approximation and the definition

$$\chi_n(k)=\int d^3y \frac{\phi_n(y)}{y}\exp(i\mathbf{k}\cdot\mathbf{y}) \tag{17}$$

we obtain

$$T_5 = iZ^{-1/2}\int dt\, d^3y \phi_n{}^*(\mathbf{y}-\mathbf{R})\frac{-2}{|\mathbf{y}-\mathbf{R}|}g_m(y) \times \exp(-\tfrac{1}{2}i\mathbf{y}\cdot\dot{\mathbf{R}}+i\delta_{mn}t)$$

$$= -(i/V)Z^{-1/2}\int \frac{d^3k}{(2\pi)^3}\exp(-i\mathbf{k}\cdot\mathbf{b})\chi_n{}^*(k)\,\delta \times\left(k_z-\frac{\delta_{mn}}{V}\right)\int d^3y g_m(y)\exp[i(\mathbf{k}-\tfrac{1}{2}\dot{\mathbf{R}})\cdot\mathbf{y}] \tag{18}$$

where g_m is the mth Hartree–Fock orbital of the ground state. Our assumptions concerning the atomic wave functions allow only ionic states which are the ground state with the mth orbital missing, i.e. single-particle excitations. This gives the amplitude for the capture of a particular electron. The proton may capture any one of the electrons and the Pauli principle gives the proper phases for combining these amplitudes. The result in this approximation is a simple addition of the amplitudes so that (18) must be multiplied by Z. In order to find the total cross section (18) must then be squared and integrated over all impact parameters. The result is

$$\frac{\sigma_{mn}}{a^2}=\frac{Z}{\pi^2V^2}\int d^3k\,|\chi_n(k)|^2\delta\left(k_z-\frac{\delta_{mn}}{V}\right)\left|\int d^3y\exp[i(\mathbf{k}-\tfrac{1}{2}\dot{\mathbf{R}})\cdot\mathbf{y}]g_m(y)\right|^2. \tag{19}$$

We now consider the high energy limit, $V\to\infty$. Then V will set the scale in the k integration so that we are interested in $k\sim V$. For large k, and s states we obtain

$$\left|\int d^3y\exp[i(\mathbf{k}-\tfrac{1}{2}\dot{\mathbf{R}})\cdot\mathbf{y}]g_m(y)\right|^2=\left[\frac{8\pi Z g_m(0)}{(\mathbf{k}-\frac{1}{2}\dot{\mathbf{R}})^4}\right]^2\left[1+O\left(\frac{Z^2}{k^2}\right)\right]. \tag{20}$$

For higher angular momentum states the leading term is at most k^{-10} which we drop. We may replace δ_{mn}/V in (19) by $\frac{1}{4}V$ by dropping terms of order Z^2/V^2 as in (20). Then (19) becomes

$$\frac{\sigma_{mn}}{a^2}=\frac{64Z^3}{V^2}|g_m(0)|^2\int\frac{d^3k}{|\mathbf{k}-\frac{1}{2}\dot{\mathbf{R}}|^8}|\chi_n(k)|^2\delta(k_z-\tfrac{1}{4}V). \tag{21}$$

Again for large k the hydrogenic s states dominate and

$$|\chi_n(k)|^2=[4\pi\phi_n(0)/k^2]^2[1+O(k^2n^2)^{-1}]. \tag{22}$$

The error here is less than in (20). The integration in (21) may now be performed and we may sum over all possible final states of the ion and hydrogen. The result is

$$\sigma/a^2=(2^{18}/5)(1.201)\pi^2(Z^3/E^6)n_A(0)[1+O(Z^2/E)] \tag{23}$$

where E is the proton laboratory energy in units of $\frac{1}{2}M(e^2/\hbar)^2=25$ keV and $n_A(0)$ is the electron density

at the origin in the atom. The factor 1.201 arises from the sum over all hydrogenic states

$$(\sum_{1}^{\infty} n^{-3} = 1.201).$$

The coefficient of the E^{-7} term depends upon further details of the atomic structure.

If we had used the form T_1, then in first Born approximation we should have obtained an additional term

$$T_1 - T_5 = \Delta_{15} = -2i(Z-1)\int dt\, d^3y_1 \cdots d^3y_Z \phi_n^*(\mathbf{y}_1 - \mathbf{R})\, u_m^{(+)*}(y_2 \cdots)$$

$$\times |\, \mathbf{y}^2 - \mathbf{R}\,|^{-1} u_0(y_1 \cdots) \exp(-\tfrac{1}{2} i \mathbf{y}_1 \cdot \dot{\mathbf{R}} + i\delta_{mn} t). \qquad (24)$$

We estimate the contribution of this term to the cross section in the high energy limit. We use

$$\int d^3y_3 \cdots d^3y_Z u_m^{(+)*}(y_2 \cdots) u_0(y_1 \cdots) \sim g_r^*(y_2)[g_a(y_1) g_b(y_2) - g_a(y_2) g_b(y_1)] \qquad (25)$$

where now the state $[m]$ of the ion is the same as the ground state with a and b deleted and r added. The calculation will show that the contribution from Δ_{15} can be dropped, so for simplicity we keep only the first term of (25). We define

$$F_{rb}(R) \sim \int d^3y_2 \frac{g_r^*(y_2) g_b(y_2)}{|\, \mathbf{y}_2 - \mathbf{R}\,|}. \qquad (26)$$

Then the contribution from Δ_{15} is similar to the first form of T_5 in (18) except that the potential $-2/|\, \mathbf{y}_1 - \mathbf{R}\,|$ is replaced by $-F_{rb}(R)$ which may be considered to be a distortion potential in the sense of (9). It has the important property that it is finite at $R=0$.

If we had used T_3 and the Born approximation, then we should obtain a result differing from T_5 by another distortion potential. It would have the form

$$v(R) = Z\int d^3x_1 \cdots d^3x_Z\, |\, u_0(x_1 \cdots)\,|^2 \frac{-2}{|\, \mathbf{x}_1 - \mathbf{R}\,|}$$

$$= -2\sum_m \int d^3x \frac{|\, g_m(x)\,|^2}{|\, \mathbf{x} - \mathbf{R}\,|} \qquad (27)$$

where the sum extends over all orbitals of the ground state. As with the previous distortion potential $v(0)$ is finite.

If we used T_6 we should also have to make an approximation for the complicated projection operator appearing there. The simplest approximation would be to replace it by the projection on to the ground state. T_6 in this approximation (and the usual Born approximation) would then become identical with the Born approximation for T_5. In the Appendix we show that the fact that the distortion potentials in these forms are all bounded at the origin is sufficient to reduce their contribution to the cross section by at least E^{-1} relative to (23). Then they give no contribution larger than E^{-7} for high energies. Indeed it would appear that they contribute in the same order as the terms dropped in (24).

It is worth pointing out here that there are previous calculations of high energy charge exchange in proton–hydrogen collisions (Jackson and Schiff 1953, Bates and Dalgarno 1952) which included the proton–proton interaction in the first Born approximation. They obtained a cross section with high energy dependence E^{-6} but with a coefficient smaller than the result obtained by omitting this interaction. In our language, the distortion potential which they included was (incorrectly) singular at $R=0$. This gives the additional E^{-6} dependence.

Thus, of the forms considered, the first Born approximation to each gives a high energy cross section behaving like E^{-6} with each form giving the same result. The coefficient of the next term (E^{-7}) would appear to depend upon the details of the atomic structure and on the form of T which is used.

We have confined our attentions to the "prior" forms of T. Evaluation of the "post" forms (T_2

and T_4) would yield a cross section with the same energy dependence as (23) but with a different coefficient. It has been proved that T_1 and T_2 are equal in the first Born approximation provided that the exact bound states are used. Thus the difference in results between T_1 and T_2 is merely a reflection of our ignorance of the exact bound states. This is not of interest here.

SECOND BORN APPROXIMATION

A classical calculation of this process has been published by Thomas (1927). He treated the problem as two successive close collisions. In the first, the slowly moving electron is struck by the proton and deflected toward the nucleus where it suffers another close collision and is deflected with essentially the same velocity as the proton. Then the relative velocity of the electron and proton must be small enough for the capture to take place. The speed of the electron does not change in the second collision so that it must acquire the proton speed in the first collision. The geometry of this situation demands two successive deflections of about 60° in the respective centre-of-mass frames. The impact parameter for such a collision between electron and proton is of the order of $(e^2/\hbar V)^2$. Thus during the collision, the ratio of the electron–proton potential to the electron–nucleus potential is $V_{pe}/V_{Ze}\sim(\hbar V/Ze^2)^2$ (the electron is assumed to be a distance a/Z from the nucleus) which is assumed large. So the approximation of a two-body collision is good in the first instance. Similar reasoning shows it to be better in the second collision. The second collision could also involve an electron instead of the nucleus.

Thomas derives the following classical results. For the first process he obtains a cross section

$$\tfrac{64}{3}\sqrt{2}\pi a^2Z^3r^{-7/2}(2/V)^{11} \qquad (28)$$

where we have inserted a factor of Z to allow for capture of any electron. Here r is the original radial position of the electron relative to the nucleus. For the second process he obtains

$$(4\pi/3)a^2Z^2r^{-7/2}(2/V)^{11}. \qquad (29)$$

Recently Drisko has carried out the second Born approximation calculation for the first process for the case of hydrogen. He treats the high energy limit and allows the electron to propagate freely in the intermediate state. From the square of the second Born approximation for T he obtains a result similar to (28) but larger by a factor $3\pi/\sqrt{2}$ if $r=1$ in (28). The interference between the second and the first Born approximation reduces the result of the first Born approximation (23) by a factor of about 0.3. Drisko has also estimated the contribution of third-order terms. He finds that there may be an E^{-6} contribution to the cross section. He points out that this may be an indication of a lack of convergence of the Born approximation series. Indeed there are strong indications (Aaron, Amado and Lee 1961) that the Born series for a rearrangement collision diverges at all energies.

If we translate Drisko's results to our problem then (23) should be multiplied by

$$1+(5\pi/2^{10})\,(E^{1/2}/Z).$$

We have not included the result analogous to (29) since the numerical coefficient is small and it has one less power of Z than (28). The modification is the result of the square of the second-order matrix element. The interference between first and second order reduces the first term in this bracket but we have seen that higher order terms contribute in this order. We shall ignore these corrections.

The coefficient of the $E^{-11/2}$ term is small here so that it is important at only extremely high energies. (It is a 10% correction at 1 meV for $Z=1$.) It is to be hoped that the correction terms in (20) which arise from overlap of atomic wave functions will have coefficients of order unity so that our calculation of an energy dependence may be expected to apply to the restricted energy region

$$10<(E/Z^2)<42.$$

This allows for a 10% correction on both ends of the energy range. For neon, $Z=10$, this is from 250 mev to about 100 mev.

ACKNOWLEDGMENTS

This work was begun while the author was a guest of the Theoretical Physics Division at the Goddard Space Flight Centre. It was finished at University College London where the author held a senior Department of Scientific and Industrial Research fellowship. Thanks are due for the hospitality and support of both groups.

APPENDIX

In Section 3 the statement was made that the contribution to the cross section from the distortion potentials was higher order in E^{-1} than (23). We demonstrate this with the interference term between this distortion potential and the leading term (18). The square of the distortion potential gives an even smaller term. The interference term is

$$Z\int d^2b\int dt\,dt'\exp[i\delta_{mn}(t-t')]\,d^3y\,d^3y'\phi_n^*(\mathbf{y}-\mathbf{R})\phi_n(\mathbf{y}'-\mathbf{R}')\frac{2}{\mathbf{y}-\mathbf{R}}\nu(\mathbf{R}')$$
$$\times g_m(y)g_m^*(y')\exp[\tfrac{1}{2}i\dot{\mathbf{R}}\cdot(\mathbf{y}-\mathbf{y}')]. \qquad \text{(A1)}$$

If we use the transforms

$$\lambda_n(k)=\int d^3y\exp(-i\mathbf{k}\cdot\mathbf{y})\phi_n(y)$$

$$\nu(l)=\int\frac{d^3R}{(2\pi)^3}\exp(-i\mathbf{l}\cdot\mathbf{R})\nu(R)$$

$$\gamma_m(k)=\int d^3y\exp(-i\mathbf{k}\cdot\mathbf{y})g_m(y)$$

then (A1) may be written

$$\frac{Z}{2\pi^2V^2}\int d^3k\,d^3l\chi_n^*(k)u_m(\mathbf{1}-\mathbf{k})\nu(l)\gamma_m(\mathbf{k}-\tfrac{1}{2}\dot{\mathbf{R}})\gamma_m(\mathbf{1}-\mathbf{k}+\tfrac{1}{2}\dot{\mathbf{R}})\delta[k_z-(\delta_{mn}/V)]. \qquad \text{(A2)}$$

We do not know the precise form of $\nu(l)$ but we do know that $\nu(R)$ is finite at $R=0$, and vanishes at $R\to\infty$. Then we may take

$$\nu(l)=-(Z/\pi^2l^2)[\beta^2/(l^2+\beta^2)]^j \qquad \text{(A3)}$$

where $j\geq 1$. Similarly, for order-of-magnitude calculation we may approximate γ_m by

$$\gamma_m(k)=\frac{8\pi Zg_m(0)}{(k^2+\alpha_m^2)^2} \qquad \text{(A4)}$$

where β is the order of the inverse of the atomic radius and α_m is a similar number.

If we now let $\delta_{mn}/V=\frac{1}{4}V$ and scale the integration variables in terms of V and let V^{-1} vanish where possible, the integral becomes

$$-Z^4\frac{|32g_m(0)|^2}{V^{13+2j}}\beta^{2j}\int\frac{d^2k\,d^2l\,dl_z}{(k^2+\frac{1}{4})^3[l^2+l_z^2][l^2+l_z^2+\beta^2/V^2]^j}$$
$$\times\frac{1}{[(\mathbf{1}-\mathbf{k})^2+(l_z-\frac{1}{4})^2+1/V^2]^2}\,\frac{1}{[(\mathbf{1}-\mathbf{k})^2+(l_z+\frac{1}{4})^2+\alpha^2/V^2]^2}. \qquad \text{(A5)}$$

We have specialized to the ground state of hydrogen. The integral has three well separated peaks at $l=l_z=0$ and $\mathbf{1}=\mathbf{k}$, $l_z=\pm\frac{1}{4}$ and we may evaluate their contributions separately in the limit of large V. At the first point the entire expression is of the order of E^{-7} which we drop. At the last pair of points it is $E^{-(6+j)}$. Since $j\geq 1$ this is also negligible.

REFERENCES

AARON, R., AMADO, R., and LEE, B., *Phys. Rev.* **121,** 319. 1961.

BATES, D. R., *Proc. Roy. Soc.* **A247,** 294, 1959.

BATES, D. R., and DALGARNO, A., *Proc. Phys. Soc.* **A65,** 919, 1952.

BATES, D. R., and MCCARROLL, *Advanc. in Phys.* **11,** No. 41, 1962.

BOHR, N., *Math.-fys. Medd.,* **18,** 8, 1940.

BRINKMAN, H. C., and KRAMERS, H. A., *Proc. Acad. Sci. Amst.* **33,** 973, 1930.

FRAME, J. W., *Proc. Camb. Phil. Soc.,* **27,** 511, 1931.

GELL-MANN, M., and GOLDBERGER, M., *Phys. Rev.* **91,** 398, 1953.

GREIDER, K., *Phys. Rev.* **114,** 786, 1959.

JACKSON, J. D., and SCHIFF, H., *Phys. Rev.,* **89,** 359, 1953.

LIPPMANN, B. A., *Phys. Rev.,* **102,** 264, 1956.

MITTLEMAN, M. H., *Phys. Rev.* **122,** 499, 1961a.

———, *Phys. Rev.,* **122,** 1930, 1961b.

———, *Phys. Rev.,* **126,** 373, 1962.

MITTLEMAN, M. H., and PU, R., *Phys. Rev.,* **126,** 370, 1962.

SCHIFF, L., *Quantum Mechanics* (New York: McGraw-Hill), 1949.

THOMAS, L. H., *Proc. Roy. Soc.* **A114,** 561, 1927.

ELECTRON IMPACT $2s$ AND $2p$ EXCITATIONS OF ATOMIC HYDROGEN

KAZEM OMIDVAR

Goddard Space Flight Center

A numerical calculation has been carried out to evaluate the 3×3 cross-section matrix involved in the electron impact excitation of the ground state of H atom to the $2s$ and the $2p$ levels. The method of solution is that of atomic eigenstates expansion. Similar calculation has been carried out by Burke, Schey, and Smith. In this paper, however, the definite integral terms in the coupled radial differential equations are eliminated through some linear transformation of the radial functions, thus avoiding iteration of these equations. As a result, the equation of reciprocity and the equation of continuity of currents are numerically satisfied with an error to value ratio less than 1 per 1000 on the average, and the maximum of this ratio, except for a few cases, has been kept below 1%. A simple perturbation theory has been developed to evaluate the effect of the long range and the centrifugal potentials. The five cross sections, $1s$–$2s$, $1s$–$2p$, $1s$–$1s$, $2s$–$2s$, and $2p$–$2p$, are tabulated.

The agreement of the $1s$–$2s$ cross sections with the experimental results are satisfactory while the calculated $1s$–$2p$ cross sections are higher than the corresponding experimental values. The $2s$–$2s$ and the $2p$–$2p$ cross sections have large values, specially at the thresholds. A Ramsauer effect in the partial cross sections of the $2s$–$2s$ elastic scattering is observed. By comparison with the eigenstate expansion calculation it is found that the Born approximation, despite its simplicity, gives meaningful results for low and close to threshold energies of the bombarding electrons. In this paper the effect of the electron spin and the exchange potentials are also investigated.

I. INTRODUCTION

The recent experimental results of Lichten and Schultz[1] on one hand, and Stebbing et al.[2] on the other, of the $1s\rightarrow2s$ transition cross section in the hydrogen atom by electron impact and the apparent discrepancy in these measurements, and the measurement of $1s\rightarrow2s$ transition by Fite et al.[3] also in hydrogen, necessitate accurate calculation for these cross sections.

Calculation of the excitation cross sections in atomic hydrogen by electron impact corresponds to the solution of the problem of three interacting bodies, one proton and two electrons. By taking the position of the proton as the center of mass, the problem will reduce to the task of finding the nonseparable wave function of the system of the two electrons with an attractive center of force. Such solution has not been found yet. However, if this wave function is expanded in terms of the eigenstates of the hydrogen atom, it is shown in this paper that through numerical integration the coefficients of the expansion, which are functions of the position vector of the free electron, can be found exactly. When an infinite number of terms are included in the expansion the solution to the problem is exact. Furthermore, the expansion has the advantage that the asymptotic form of its coefficients are automatically the asymptotic form of the free electron wave function scattered from different atomic states, which are simply related to the excitation cross sections.

In this paper atomic states $1s$, $2s$, $2p$ are included in the expansion, and by antisymmetrizing the two electron wave functions according to the exclusion principle, some contribution from the continuum in the expansion is also taken into account. The first calculation of this type was performed by Marriot[4] whose expansion consisted of the $1s$ and the $2s$ states in order to calculate the $1s\rightarrow2s$ transition cross section. This calcula-

tion was extended by Smith[5] to higher total orbital angular momenta of the system. Percival and Seaton[6] have formulated the eigen state expansion technique in general, and have tabulated the coefficients of the integro-differential equations for s, p, and d atomic electrons. While this paper was in preparation, Burke, Smith, and Shey,[7,32] using the tables of reference 6 for three states $1s$, $2s$, $2p$, have integrated the resulting integro-differential equations. Their technique of solution is substantially different from that of this paper in more than one respect, and it is believed that the results presented here are more accurate.

The degeneracy of the $2s$ and the $2p$ levels of the hydrogen atom makes it necessary, as pointed out by Massey,[8] that any calculation concerning $1s\rightarrow 2s$ or $1s\rightarrow 2p$ transitions contain these two states simultaneously. In this respect the present calculation and that of reference 7 is superior to that of reference 4, where $1s$ and $2s$ states are included to calculate $1s\rightarrow 2s$ transition, or that of Khashaba and Massey,[9] where $1s$ and $2p$ states are included to calculate $1s\rightarrow 2p$ transition.

We treat the present problem in the following manner. In Sec. IIA we derive the integro-differential equations of the problem independent of reference 6. Sec. IIB is the derivation of the transmission matrix and the cross sections. In Sec. IIC we derive a useful relationship for numerical integration based on the symmetry of the interacting potentials. Sec. IID is the derivation of the partial wave Born approximation transmission matrix. Different parts in Sec. III deal with techniques of numerical integration and the effects of the long range potentials. The transformation of the radial wave functions presented here avoid the need for iteration of the coupled integro-differential equations, as is the case in reference 7. Finally, Sec. IV gives the results of numerical integration and their interpretation.

The numerical integrations were carried out for all partial waves, where in higher partial waves Born approximation were used. The transition between the eigenstate expansion calculation and Born approximation takes place when the results of the two calculations agree closely.

II. FORMULATION

A. Derivation of the Differential Equations

Since spin orbit interaction of the electrons are neglected, the total orbital angular momentum L, and the total spin angular momentum S, are separately conserved. We can then divide the interactions into antiparallel spin states where $S=0$, and parallel spin states where $S=1$. We then deal with spatial wave functions of the electrons only and for brevity we call the orbital angular momentum the angular momentum.

Neglecting the motion of the proton of the hydrogen atom and taking its position as the origin of the coordinate system, the Schroedinger equation for the system can be written

$$[H-E]\psi(r_1, r_2)=0, \tag{2.1}$$

where r_1 and r_2 are the position vectors of the bound and free electrons, and in atomic units

$$H-E=-\tfrac{1}{2}\nabla_1^2-\tfrac{1}{2}\nabla_2^2-r_1^{-1}-r_2^{-1}+r_{12}^{-1}-E, \tag{2.2}$$

where E is the total energy of the system and r_{12} is the distance between the two electrons. We expand the total wave function $\psi(r_1, r_2)$ in terms of the eigenfunctions of the total angular momentum L,

$$\psi(r_1, r_2)=\sum_{L=0}^{\infty}\psi_L(r_1, r_2). \tag{2.3}$$

Since these eigenfunctions are orthogonal and distinct, substitution of Eq. (2.3) in Eq. (2.1) gives,

$$[H-E]\psi_L(r_1, r_2)=0. \tag{2.4}$$

The explicit form of $\psi_L(r_1, r_2)$ is given by

$$\psi_L(r_1, r_2)=(1+\beta P_{12})\sum_{n_1l_1l_2}\sum_{m_1m_2}C_{m_1m_2M}^{l_1l_2L}\varphi(n_1l_1m_1, r_1)r_2^{-1}u(k_{n_1}l_2, r_2)\,Y_{l_2m_2}(\Omega_2), \tag{2.5}$$

$$\varphi(n_1l_1m_1, r_1)=r_1^{-1}P(n_1l_1, r_1)\,Y_{l_1m_1}(\Omega_1). \tag{2.6}$$

Here $\varphi(n_1l_1m_1, r_1)$ is the hydrogen atom wave function with radial part $r_1^{-1}P(n_1l_1, r_1)$ and angular part $Y_{l_1m_1}(\Omega_1)$ and quantum numbers nl_1m_1; $r_2^{-1}u(k_{n_1}l_2, r_2)$ is the radial part and $Y_{l_2m_2}(\Omega_2)$ is the angular

part of the free electron wave function with quantum numbers $k_{n_1} l_2 m_2$. The relationship between the wave number k_{n_1} and n_1 is given by

$$k_{n_1}{}^2 = 2[E + (1/2n_1{}^2)]. \tag{2.7}$$

Finally the constants $C_{m_1 m_2 M}{}^{l_1 l_2 L} = (l_1 l_2 m_1 m_2 \mid LM)$ are vector coupling coefficients which make the linear combination of the products of the one electron wave functions in Eq. (2.5) the eigenfunction of L. In the problem under consideration $n_1 = 1, 2$; $l_1 = 0, 1$; $l_2 = \mid L - l_1 \mid, \cdots, \mid L + l_1 \mid$; $m_1 = -l_1, \cdots, l_1$, and $m_2 = -l_2, \cdots, l_2$. To make the total wave function symmetric for antiparallel spins or antisymmetric for parallel spins, the operator P_{12} interchanges r_1 and r_2 while β is $+1$ for the first case and is -1 for the second.

By taking L along the z-axis, $M = 0$ and $m_2 = -m_1$. Equation (2.5) can then be written

$$\psi_L(r_1, r_2) = (1 + \beta P_{12}) \sum_{n_1 l_1 l_2} \sum_{m_1} C_{m_1 - m_1 0}{}^{l_1 l_2 L} \varphi(n_1 l_1 m_1, r_1) r_2{}^{-1} u(k_{n_1} l_2, r_2)\, Y_{l_2 m_2}(\Omega_2). \tag{2.8}$$

In order that $\psi_L(r_1, r_2)$ closely approximates the exact wave function, we minimize the expectation value of the energy operator with respect to the radial parts of the free electron wave functions,

$$\delta \int \psi_L{}^*(r_1, r_2)[H - E]\psi_L(r_1, r_2)\, d^3 r_1\, d^3 r_2 = 0. \tag{2.9}$$

It has been shown by Kohn[10] that the differences between the scattering amplitudes obtained from these equations and the exact scattering amplitudes are quadratic in the difference between $\psi_L(r_1, r_2)$ and the exact wave function. When the variation is carried out inside the integral we obtain

$$\sum_{m_1} C_{m_1 - m_1 0}{}^{l_1 l_2 L} \int \varphi^*(n_1 l_1 m_1, r_1)\, Y_{l_2 m_2}{}^*(\Omega_2)[H - E]\psi_L(r_1, r_2)\, d^3 r_1\, d\Omega_2 = 0. \tag{2.10}$$

By means of Eqs. (2.2), (2.8), the Schroedinger equation for the hydrogen atom,

$$[\, \nabla_1{}^2 + (2/r_1)\,]\varphi(n_1 l_1 m_1, r_1) = (n_1{}^2)^{-1} \varphi(n_1 l_1 m_1, r_1), \tag{2.11}$$

and Eq. (2.7), Eq. (2.10) reduces to

$$\sum_{m_1} \sum_{n_1' l_1' l_2'} \sum_{m_1'} C_{m_1 - m_1 0}{}^{l_1 l_2 L} C_{m_1' - m_1' 0}{}^{l_1' l_2' L} \int \varphi^*(n_1 l_1 m_1, r_1)\, Y_{l_2 m_2}{}^*(\Omega_2)\,(1 + \beta P_{12})$$

$$\times \left[\nabla_{r_2}{}^2 - \frac{l_2'(l_2' + 1)}{r_2{}^2} + k_{n_1'}{}^2 + 2(r_2{}^{-1} - r_{12}{}^{-1}) \right] \varphi(n_1' l_1' m_1', r_1) r_2{}^{-1} u(k_{n_1'} l_2', r_2)\, Y_{l_2' m_2'}(\Omega_2)\, d^3 r_1\, d\Omega_2 = 0, \tag{2.12}$$

where $\nabla_{r_2}{}^2$ is the radial part of $\nabla_2{}^2$. By orthogonality of the hydrogen atom and spherical harmonics wave functions, the relation[11]

$$\sum_{m_1} [C_{m_1 - m_1 0}{}^{l_1 l_2 L}]^2 = 1, \tag{2.13}$$

the integration by parts of the exchange terms, and the relation[12]

$$C_{-m_1 m_1 0}{}^{l_2 l_1 L} = (-)^{L - l_1 - l_2} C_{m_1 - m_1 0}{}^{l_1 l_2 L}, \tag{2.14}$$

Eq. (2.12) leads to

$$\left[\nabla_{r_2}^2-\frac{l_2(l_2+1)}{r_2^2}+k_{n_1}^2+\frac{2}{r_2}\right]\frac{u(k_{n_1}l_2,r_2)}{r_2}-2\sum_{m_1}\sum_{n_1'l_1'l_2'}\sum_{m_1'}C_{m_1-m_10}^{l_1l_2L}C_{m_1'-m_1'0}^{l_1'l_2'L}$$

$$\times\int\frac{\varphi^*(n_1l_1m_1,r_1)\,Y^*_{l_2m_2}(\Omega_2)}{r_{12}}\Bigg[\varphi(n_1'l_1'm_1',r_1)\frac{u(k_{n_1}'l_2',r_2)}{r_2}Y_{l_2'm_2'}(\Omega_2)+\beta\varphi(n_1'l_1'm_1',r_2)$$

$$\times\frac{u(k_{n_1}'l_2',r_1)}{r_1}Y_{l_2'm_2'}(\Omega_1)\Bigg]d^3r_1\,d\Omega_2+\beta\sum_{n_1'l_1'l_2'}(-)^{L-l_1-l_2}\delta(l_1'l_2',l_2l_1)\left[(n_1^2)^{-1}+k_{n1'}^2\right]$$

$$\times\int_0^\infty P(n_1l_1,r_1)r_2^{-1}P(n_1'l_1',r_2)u(k_{n_1'}l_2',r_1)\,dr_1=0. \qquad (2.1\,)$$

If $1/r_{12}$ is expanded in terms of the Legendre polynomials and use is made of the addition theorem[13] we obtain

$$1/r_{12}=\sum_{\lambda=0}^{\infty}(r_<^\lambda/r_>^{\lambda+1})\,P_\lambda(\cos\theta_{12})$$

$$=\sum_{\lambda=0}^{\infty}\sum_{\mu=-\lambda}^{+\lambda}\left[4\pi/(2\lambda+1)\right](r_<^\lambda/r_>^{\lambda+1})\,Y_{\lambda\mu}(\Omega_1)\,Y^*_{\lambda\mu}(\Omega_2). \qquad (2.16)$$

In this expression θ_{12} is the angle between the position vectors r_1 and r_2 at the origin, and $r_<$ is the smaller and $r_>$ is the larger of $|\,r_1\,|$ and $|\,r_2\,|$. We also introduce

$$y_\lambda(nln'l',r_2)=r_2^{-(\lambda+1)}\int_0^{r_2}P(nl,r_1)\,P(n'l',r_1)r_1^\lambda\,dr_1+r_2^\lambda\int_{r_2}^\infty P(nl,r_1)\,P(n'l',r_1)r_1^{-(\lambda+1)}\,dr_1. \qquad (2.17)$$

Then it follows that

$$\int_0^\infty\frac{P(nl,r_1)\,P(n'l',r_1)}{r_{12}}\,dr_1=\sum_{\lambda=0}^{\infty}\sum_{\mu=-\lambda}^{+\lambda}Y_{\lambda\mu}(\Omega_1)\,Y^*_{\lambda\mu}(\Omega_2)y_\lambda(nln'l',r_2). \qquad (2.18)$$

By means of Eq. (2.18), the relation[14]

$$\int Y^*_{l_3m_3}Y_{l_2m_2}Y_{l_1m_1}\,d\Omega=\left[\frac{(2l_1+1)\,(2l_2+1)}{4\pi(2l_3+1)}\right]^{1/2}C_{m_1m_2m_3}^{l_1l_2l};C_{000}^{l_1l_2l_3}, \qquad (2.19)$$

and the definition

$$(nl\mid k_{n_1}l')=\int_0^\infty P(nl,r)u(k_{n_1}l',r)\,dr, \qquad (2.20)$$

Eq. (2.15) when multiplied by r_2 gives

$$\left[\frac{d^2}{dr_2^2}-\frac{l_2(l_2+1)}{r_2^2}+k_{n_1}^2+\frac{2}{r_2}\right]u(k_{n_1}l_2,r_2)-2\left(\frac{2l_2+1}{2l_1+1}\right)^{1/2}\sum_{n_1'l_1'l_2'}\sum_{m_1m_1'}\sum_{\lambda=0}^{\infty}\sum_{\mu=-\lambda}^{+\lambda}C_{m_1-m_10}^{l_1l_2L}C_{m_1'-m_1'0}^{l_1'l_2'L}$$

$$\times\Bigg\{\left(\frac{2l_1'+1}{2l_2'+1}\right)^{1/2}C_{\mu m_1'm_1}^{\lambda l_1'l_1}C_{000}^{\lambda l_1'l_1}C_{\mu m_2m_2'}^{\lambda l_2l_2'}C_{000}^{\lambda l_2l_2'}y_\lambda(n_1l_1n_1'l_1',r_2)u(k_{n_1'}l_2',r_2)$$

$$+\beta\left(\frac{2l_2'+1}{2l_1'+1}\right)^{1/2}C_{\mu m_2'm_1}^{\lambda l_2'l_1}C_{000}^{\lambda l_2'l_1}C_{\mu m_2m_1'}^{\lambda l_2l_1'}C_{000}^{\lambda l_2l_1'}P(n_1'l_1',r_2)y_\lambda(n_1l_1k_{n_1'}l_2',r_2)\Bigg\}$$

$$+\beta\sum_{n_1'l_1'l_2'}(-)^{L-l_1-l_2}\delta(l_1'l_2',l_2l_1)\left[(n_1^2)^{-1}+k_{n_1'}^2\right)P(n_1'l_1',r_2)\,(n_1l_1\mid k_{n_1'}l_2')=0. \qquad (2.21)$$

In the exchange integrals above we have defined $P(k_{n_1}l_2, r)$ as $u(k_{n_1}l_2, r)$.

The vectorial equations

$$L=l_1+l_2, \tag{2.22}$$

where L is constant but l_1 and l_2 take the values given before, can be divided into two groups, one with $L-l_1-l_2$ even and the other with $L-l_1-l_2$ odd. Since the total spacial wave function has the parity of l_1+l_2, in the first group the wave function has the parity of L and in the second a parity opposite to L. By conservation of parity we have two distinct groups of interactions. In this problem, where $1s$, $2s$ and $2p$ states of atomic hydrogen are taken into account, it is easy to see that when $L-l_1-l_2$ is even, the set of quantum numbers $k_{n_1}l_2$ has 4 values; one for each of the $1s$ and $2s$ states and two for the $2p$ state. When $L-l_1-l_2$ is odd, $k_{n_1}l_2$ has one value which corresponds to the elastic scattering of electrons by the $2p$ state of the hydrogen atom. Equation (2.21) is evaluated for these cases and the resulting differential equations are listed in Appendix I. In evaluating Eq. (2.21) it should be noted that the $C_{m_1m_2m_3}{}^{l_1l_2l_3}$ coefficients are subject to the condition that $l_1l_2l_3$ form a closed triangle and $m_3=m_1+m_2$.[15] This limits the summation over λ and μ considerably to few terms only. Summation over m_1, m_1', λ and u are carried out using the numerical values of the C coefficients given by Condon and Shortley.[16]

Percival and Seaton[6] have derived the same differential equations for the scattering of free electrons by atomic s, p, and d electrons in hydrogen atom using the theory of irreducible tensor operators to evaluate the interaction terms between the two electrons in the differential equations. The calculation became considerably simpler in this way. The results of the two methods are identical.

In the rest of the paper except Sec. IIE we discuss the solutions to the four coupled differential equations given in Appendix I and which arise when $L-l_1-l_2$ is even. The single differential equations for $L-l_1-l_2$ odd is derived in Sec. IIE. Its numerical integration can be treated as a special case of the four coupled differential equations.

When the integrals representing the direct potentials in the four differential equations are evaluated and some change is made in the limits of the exchange potential integrals, these equations can be written in the following matrix form

$$\left[\frac{d^2}{dr^2}+k_n{}^2-\frac{l_n(l_n+1)}{r^2}\right]\mathbf{u}(k_nl_n, r)=2\mathbf{V}\mathbf{u}(k_nl_n, r). \tag{2.23}$$

The four components of $\mathbf{u}$ are the four radial functions of the free electron. $\mathbf{V}$ is a 4×4 symmetric matrix that is the sum of three matrices,

$$V_{ij}=D_{ij}+E_{ij},$$

$$E_{ij}=F_{ij}+\sum_{\nu=1}^{\sigma}g_{ij}{}^{\nu}\int_0^{\infty}h_{ij}{}^{\nu}\,dr, \tag{2.24}$$

where D_{ij} is the direct and E_{ij} is the exchange potentials and both are functions of r. The matrix E_{ij} contains in addition integrals with respect to r and for the purpose of numerical integration it can be written as the sum of two matrices. The explicit forms of D_{ij}, F_{ij}, $g_{ij}{}^{\nu}$ and $h_{ij}{}^{\nu}$ are given in Appendix II. The value of σ is 2 for $i=j=3$ and $i=j=4$, and is 1 for all other values of i and j. It is understood that for the exchange terms the components of $\mathbf{u}$ on the right-hand side of Eq. (2.23) are inside the integrals of the exchange terms.

B. Derivation of the Transmission Matrix from Solutions of the Differential Equations

The method is similar to that used by Bransden and McKee,[17] and by Marriot.[4] Equation (2.23) constitutes a set of four coupled, second-order, differential equations. Three components of $\mathbf{u}$ can be eliminated from these equations, resulting in an eighth-order differential equation for the remaining component. Therefore, there are eight sets of solutions to Eq. (2.23). However, only half of these solutions are regular at the origin. Each of the four regular solutions corresponds to a definite vector $\mathbf{u}$. The four vectors can properly be represented by a 4×4 matrix u_{nj}, $n, j=1, 2, 3, 4$, where n corresponds to the particular component and j corresponds to the particular solution of $\mathbf{u}$. The four solutions are carried out numerically in the next section.

From the explicit form of **V** it can be seen that **V** vanishes at infinity. The asymptotic solutions of **u** as given by Eq. (2.23) is, therefore,

$$u_{nj}(r)\sim a_{nj}\sin[k_n r-\tfrac{1}{2}(l_n\pi)+\delta_{nj}]$$
$$n,j=1,2,3,4, \quad (2.25)$$

where a_{nj} is the amplitude and δ_{nj} is the phase shift of the jth solution of the nth component of **u**.

Corresponding to the four components of **u** there are four channels open to the reaction. If the incident wave be in the mth channel ($m=1$, 2, 3, 4), the travelling wave in the n channel will be given by

$$u_n(r)\sim \exp[-i(k_n r-\tfrac{1}{2}l_n\pi)]\delta(m,n)-S_{mn}\exp[i(k_n r-\tfrac{1}{2}l_n\pi)] \qquad n=1,2,3,4. \quad (2.26)$$

The constants S_{mn} are the amplitudes of the scattered waves. Since Eqs. (2.26) are also the asymptotic solutions of Eqs. (2.23), they must be equal to linear combinations of Eqs. (2.25). If we call the coefficients of the linear combinations P_j we must have

$$\sum_{j=1}^{4}P_j a_{nj}\sin[k_n r-\tfrac{1}{2}(l_n\pi)+\delta_{nj}]=(k_n)^{-1/2}\{\exp[-i(k_n r-\tfrac{1}{2}l_n\pi)]\delta(m,n)-S_{mn}\exp[i(k_n r-\tfrac{1}{2}l_n\pi)]\}$$
$$n,m=1,2,3,4. \quad (2.27)$$

On the right-hand side, we have used the normalization of Blatt and Weisskopf.[18] If we equate the coefficients of

$$\exp[-i(k_n r-\tfrac{1}{2}l_n\pi)]$$

and

$$\exp[i(k_n r-\tfrac{1}{2}l_n\pi)]$$

in Eqs. (2.27) we obtain

$$\sum_{j=1}^{4}P_j a_{nj}\exp-[i\delta_{nj}]=[-2i/(k_n)^{1/2}]\delta(m,n)$$
$$\sum_{j=1}^{4}P_j a_{nj}\exp[i\delta_{nj}]=[-2i/(k_n)^{1/2}]S_{mn} \quad (2.28)$$

Separation of Eqs. (2.28) into real and imaginary parts gives

$$\sum_{j=1}^{4}[(\Re P_j)\sin\delta_{nj}-(\Im P_j)\cos\delta_{nj}]a_{nj}=[2/(k_n)^{1/2}]\delta(m,n)$$
$$\sum_{j=1}^{4}[(\Re P_j)\cos\delta_{nj}+(\Im P_j)\sin\delta_{nj}]a_{nj}=0$$
$$\sum_{j=1}^{4}[(\Re P_j)\sin\delta_{nj}+(\Im P_j)\cos\delta_{nj}]a_{nj}=[-2/(k_n)^{1/2}]\Re S_{mn}$$
$$\sum_{j=1}^{4}[(\Re P_j)\cos\delta_{nj}-(\Im P_j)\sin\delta_{nj}]a_{nj}=[2/(k_n)^{1/2}]\Im S_{mn} \quad (2.29)$$

In the above $\Re$ or $\Im$ represent the real or the imaginary part of the quantity that follows then. Equations (2.29) are a set of 16 linear equations for 16 unknowns $\Re P_j$, $\Im P_j$, $\Re S_{mn}$ and $\Im S_{mn}$. Once these unknowns are found,[19] the magnitude of S_{mn} will be given by

$$|S_{mn}|^2=(\Re S_{mn})^2+(\Im S_{mn})^2. \quad (2.30)$$

The cross section is obtained by asymptotic expansion in spherical harmonics of the incident plane wave[20]

$$\exp[ikz]\sim(\pi^{1/2}/kr)\sum_{l=0}^{\infty}(2l+1)^{1/2}i^{l+1}\{\exp[-i(kr-\tfrac{1}{2}l\pi)]-\exp[i(kr-\tfrac{1}{2}l\pi)]\}Y_{l,0}. \quad (2.31)$$

The magnitude of the ingoing wave on the right-hand side of Eqs. (2.27) for $n=m$ is $[k_m/\pi(2l_m+1)]^{1/2}$ times the magnitude of the partial wave of the expansion of $r\exp[ik_mz]$. The plane wave has a flux of v which, in atomic units, is equal to k. The ignoing flux of the right-hand side of Eqs. (2.27) is, therefore, $k_m^2/[\pi(2l_m+1)]$. The outgoing flux in the channel $n\neq m$ is $|S_{mn}|^2$. The cross section is obtained when we average the ratio of the outgoing flux to the ingoing flux over the initial states, and sum over the final final states. For a particular spin state of the two electrons, unpolarized electron beam and unoriented atoms, the multiplicity of the initial states is $(2l_1+1)(2l_2+1)$, where l_1 and l_2 are the angular momentum of the bound and free electrons. For a polarized beam, $m_2=0$, where m_2 is the magnetic quantum number of the free electron. Then $m_1=M$, where m_1 and M are the bound electron and the total magnetic quantum numbers. Since M is constant, there is only one initial state for a polarized beam. The multiplicity of the final states is $2L+1$, where L is the total orbital angular momentum. Since $l_m=l_2$, the cross section for $m\neq n$ is

$$Q_{mn}=\frac{\pi(2L+1)}{k_m^2(2l_1+1)}|S_{mn}|^2, \qquad m\neq n. \quad (2.32)$$

The outgoing partial wave in the incident channel m consists of the scattered wave plus the outgoing wave given in the expansion of the plane wave. Then, according to Eqs. (2.27), for $n=m$ the magnitude of the amplitude of the scattered wave is $|1-S_{mn}|$. The elastic scattering cross section is therefore given by

$$Q_{mm}=\frac{\pi(2L+1)}{k_m^2(2l_1+1)}|1-S_{mm}|^2 \quad (2.33)$$

If we define a matrix T by the relation

$$T=1-S, \quad (2.34)$$

Eqs. (2.32) and (2.33) can then be combined into a single equation,

$$Q_{mn}=\frac{\pi(2L+1)}{k_m^2(2l_1+1)}|T_{mn}|^2. \quad (2.35)$$

T_{mn} is the transmitted amplitude in the n channel due to an incident wave in the m channel. The elements of T_{mn} constitute the transmission matrix.

The matrix S has two properties that are useful as tests on the accuracy of numerical integration. From Eq. (2.26) it can be seen that S transforms the ingoing wave into the outgoing waves. The continuity of the electronic current requires that S be a unitary matrix

$$\sum_{n=1}^{4}|S_{mn}|^2=1, \qquad m=1,2,3,4. \quad (2.36)$$

Furthermore, since the Hamiltonian is Hermetian, S must be symmetric,[18]

$$S_{mn}=S_{nm}. \quad (2.37)$$

Equations (2.36) and (2.37) are used as tests on the accuracy of numerical integration.

C. A Useful Relationship

A relationship based on the symmetry of the interaction potentials can be derived which serves as another test on the accuracy of the solutions. The lth and the kth solutions of the ith component of $\mathbf{u}$ by Eq. (2.23) are given by

$$\{(d^2/dr^2)+k_i^2-l_i[(l_i+1)/r^2]\}u_{il}=\sum_j V_{ij}u_{jl}$$

$$\{(d^2/dr^2)+k_i^2-l_i[(l_i+1)/r^2]\}u_{ik}=\sum_j V_{ij}u_{jk} \quad (2.38)$$

Multiplying the first by u_{ik} and the second by u_{il}, subtracting the two expressions, and summing

over i gives

$$\sum_i [u_{ik}(d^2/dr^2)u_{il}-u_{il}(d^2/dr^2)u_{ik}] = \sum_{i,j} V_{ij}[u_{ik}u_{jl}-u_{il}u_{jk}] \quad (2.39)$$

Since $V_{ij}=V_{ji}$, the interchange of the summation indices changes the sign on the right-hand side of the equation, the right-hand side must therefore be zero. Integrating the left-hand side from zero to infinity, we obtain

$$\sum_i \int_0^\infty [u_{ik}(d^2/dr^2)u_{il}-u_{il}(d^2/dr^2)u_{ik}]\,dr=0. \quad (2.40)$$

Integrating the above equation by parts, and applying Eq. (2.25) we obtain

$$\sum_{i=1}^{4} k_i a_{ik}a_{il}\sin(\delta_{ik}-\delta_{il})=0 \qquad k,l=1,2,3,4,\ k\neq l. \quad (2.41)$$

Although the terms containing the exchange potentials do not cancel out on the right-hand side of Eq. (2.39), the cancellation does take place after the integration is carried out in Eq. (2.40).

D. Transmission Matrix According to Born Approximation

The Born approximation consists of neglecting the exchange potential terms appearing in the **V** matrix of Eq. (2.23), and also of neglecting all the direct potential terms in this matrix except those terms that connect the incident channel to all other channels.[21] Equation (2.23), when the incident wave is in the mth channel, reduces to

$$\{(d^2/dr^2)+k_n^2-[l_n(l_n+1)/r^2]\}u_n=2D_{nm}u_m, \qquad n=1,2,3,4, \quad (2.42)$$

u_m and u_n are given asymptotically by

$$u_m\sim k_m^{-1/2}\sin[k_m r-\tfrac{1}{2}(l_m\pi)] \quad (2.43)$$

$$u_n\sim k_n^{-1/2}B_{nm}\cos[k_n r-\tfrac{1}{2}(l_n\pi)]. \quad (2.44)$$

We have chosen the constants of proportionality of u_m and u_n such that B_{nm} is the Born approximation of the reactance matrix R.[22,23] Equation (2.43) shows that u_m must have the following form[24]

$$u_m=k_m^{1/2}r j_{l_m}(k_m r). \quad (2.45)$$

where $j_{l_m}(k_m r)$ are spherical Bessel functions. Furthermore, if y_n represents the homogeneous solution of Eq. (2.42), it must have the following forms

$$y_n=a_n k_n r j_{l_n}(k_n r), \quad (2.46)$$

$$y_n\sim a_n\sin[k_n r-\tfrac{1}{2}(l_n\pi/2)], \quad (2.47)$$

with a_n some unknown constant. Multiplying Eq. (2.42) on the left by y_n and integrating the result from zero to infinity we obtain by partial integration

$$\begin{aligned}2\int_0^\infty y_n D_{nm}u_m\,dr &= \int_0^\infty y_n\left[\frac{d^2}{dr^2}+k_n^2-\frac{l_n(l_n+1)}{r^2}\right]u_n\,dr\\ &=[y_n(d/dr)u_n-u_n(d/dr)y_n]_0^\infty\\ &=-k_n^{1/2}a_n B_{nm}.\end{aligned}$$

The last equality has been obtained by noticing that y_n and u_n vanish at the origin, and by using their asymptotic forms as given by Eqs. (2.44) and (2.47). We therefore have

$$B_{nm}=-2(k_n k_m)^{1/2}\int_0^\infty j_{l_n}(k_n r)D_{nm}j_{l_m}(k_m r)r^2\,dr. \quad (2.48)$$

This is identical to the expression given for **B** by Seaton.[25]

The transmission and the reactance matrices are related by $\mathbf{T}=-2i\mathbf{R}/(1-i\mathbf{R})$. Since $\mathbf{B}\ll 1$, the transmission matrix according to the Born approximation is given by

$$T_{nm}{}^B=4i(k_n k_m)^{1/2}\int_0^\infty j_{l_n}(k_n r)D_{nm}j_{l_m}(k_m r)r^2\,dr. \quad (2.49)$$

Substitution of Eq. (2.49) in Eq. (2.35) would give the cross section according to the Born approximation. It should be noted that the sym-

metry of **T** insures Eq. (2.37) to be satisfied, while Eq. (2.36) is not satisfied anymore. If we define

$$\mathbf{T}=-2i\mathbf{B}/(1-i\mathbf{B}),$$

then

$$\mathbf{S}=(1+i\mathbf{B})/(1-i\mathbf{B})$$

and the symmetry of **B** makes **S** unitary. Equation (2.36), the continuity of current, is then satisfied. This is the definition of approximation II of the Born approximation according to Seaton,[23] which will be discussed later.

E. Elastic Scattering of Electrons by the 2p States of the Hydrogen Atom

The angular momentum of the free electron, l_2, in the $2p$ channel has the values $L-1$, L, $L+1$, where L is the total angular momentum of the system. The first and the last values were considered in previous sections. The case $l_2=L$ corresponds to a wave function in the $2p$ channel with a parity different from all channel wave functions considered previously. It therefore corresponds to elastic scattering. The wave function in this case is given by

$$\psi_L(r_1, r_2)=(1+\beta P_{12})\sum_{m_1=-1}^{+1} C_{m_1-m_10}{}^{1LL}\varphi_{2pm_1}(r_1)\frac{u(k_2L, r)}{r_2}Y_{L-m_1}(\Omega_2). \tag{2.50}$$

When Eq. (2.9) is formed with this wave function, and minimized with respect to $u(k_2L, r)$, treatment which led to the derivation of the four differential equations will give the following differential equation

$$\{(d^2/dr^2)+k_2^2-[L(L+1)/r^2]+(2/r)\}u_L(r)-\beta(\tfrac{1}{4}+k_2^2)\,\delta(L, 1)rR_{21}(r)\,(2p\mid k_2L)$$
$$+2[y_0(2p2p, r)-\tfrac{1}{5}y_2(2p2p, r)]u_L(r)+2\beta rR_{21}(r)\left[-\frac{3y_{L-1}(2pk_2L, r)}{(2L+1)(2L-1)}+\frac{3y_{L+1}(2pk_2L, r)}{(2L+1)(2L+3)}\right]. \tag{2.51}$$

The asymptotic solution of this equation is given by

$$u_L\sim a_L\sin[k_2r-\tfrac{1}{2}(L\pi)+\delta_L]. \tag{2.52}$$

If the scattering amplitude is designated by T_{55}, it can easily be shown from Sec. IIB that for a particular L

$$T_{55}=1-\exp(2i\delta)=-2i\exp(i\delta)\sin\delta, \tag{2.53}$$

where, for simplicity, we have suppressed the subscript L. The cross section, according to Eq. (2.35), is given by

$$Q_{55}=\frac{4\pi(2L+1)}{3k_2^2}\sin^2\delta \tag{2.54}$$

The total elastic scattering cross section by the $2p$ states is the sum of this cross section and the cross section corresponding to $l_2=L-1$ and $l_2=L+1$, which were considered previously.

The Born amplitude, Eq. (2.48), in this case is given by

$$B_{55}=-2k_2\int_0^\infty j_L(k_2r)D_{55}j_L(k_2r)r^2\,dr, \tag{2.55}$$

where by Eq. (2.51),

$$D_{55}=-r^{-1}+y_0(2p2p, r)-\tfrac{1}{5}y_2(2p2p, r). \tag{2.56}$$

III. NUMERICAL INTEGRATION

A. Decomposition of the Differential Equations

If it were not for the definite integrals appearing in the potential matrix **V**, the set of the four coupled differential equations (2.23) could be integrated by any standard technique. The presence of these unknown constants whose integrand involve the unknown functions makes it necessary to solve these equations by iteration or by transformation of **u** into other vectors, whose differential equations do not contain definite integrals. Since the terms containing definite integrals are small as compared to the direct potentials, the iteration method can be used by assuming that the values of these integrals are zero. The differential equations are then integrated, the values of the definite integrals that are subsequently obtained are substituted in the differential equations, and the integration is repeated. The process

is repeated until sufficiently consistent values of these integrals are obtained. This method is useful if the convergences of the constants are fast enough, and the cross section is not very sensitive to the values of these constants.

In the second method, the transformation of **u** fixes the values of the constants and thus avoids iteration, whereby the computation is reduced considerably. The description of the method will given here.[26,27]

By making use of Eq. (2.24), Eq. (2.23) can be written

$$\{(d^2/dr^2)+k_i^2-[l_i(l_i+1)/r^2]\}u_i = 2\sum_{j=1}^{4}[(D_{ij}+F_{ij})u_j+\sum_{\mu=1}^{\sigma}g_{ij}{}^{\mu}C_{ij}{}^{\mu}] \quad (3.1)$$

where

$$C_{ij}{}^{\mu}=\int_0^{\infty}h_{ij}{}^{\mu}(r)u_j(r)\,dr. \quad (3.2)$$

We introduce the functions v_i and $u_i{}^{kl}$ that are solutions of the following differential equations

$$\{(d^2/dr^2)+k_i^2-[l_i(l_i+1)/r^2]\}v_i=2\sum_{j=1}^{4}[D_{ij}+F_{ij}]v_j \quad (3.3)$$

$$\{(d^2/dr^2)+k_i^2-[l_i(l_i+1)/r^2]\}u_i{}^{kl}=2\sum_{j=1}^{4}[D_{ij}+F_{ij}]u_j{}^{kl}+2\delta(i,k)g_{kl}{}^{\nu}. \quad (3.4)$$

Then, u_i is given by the following expression

$$u_i=v_i+\sum_{k=1}^{4}\sum_{l=1}^{4}\sum_{\nu=1}^{\sigma}C_{kl}{}^{\nu}u_i{}^{kl}. \quad (3.5)$$

Equation (3.5) can be verified by multiplying Eq. (3.4) by $C_{kl}{}^{\nu}$, summing over k, l, and ν, and adding to Eq. (3.3), whereupon Eq. (3.1) results. Substitution of Eq. (3.5) in Eq. (3.2) gives

$$\sum_{k=1}^{4}\sum_{l=1}^{4}\sum_{\nu=1}^{\sigma}[\delta(ij\mu,kl\nu)-A_{ij}{}^{\mu kl}]C_{kl}{}^{\nu}=B_{ij}{}^{\mu}, \qquad i,j=1,2,3,4;$$

$$\mu=1,2 \quad \text{for} \quad i=j=3 \quad \text{and} \quad i=j=4; \qquad \mu=1 \quad \text{otherwise}, \quad (3.6)$$

where $A_{ij}{}^{\mu kl}$ and $B_{ij}{}^{\mu}$ are defined by

$$A_{ij}{}^{\mu kl}=\int_0^{\infty}h_{ij}{}^{\mu}u_j{}^{kl}\,dr$$

$$B_{ij}{}^{\mu}=\int_0^{\infty}h_{ij}{}^{\mu}v_j\,dr \quad (3.7)$$

The numerical integration is carried out by integrating Eqs. (3.3) and (3.4) by any standard method, calculating $A_{ij}{}^{\mu kl}$ and $B_{ij}{}^{\mu}$ by Eqs. (3.7) and, finally, solving the system of 18 algebraic equations given by Eqs. (3.6) to find $C_{kl}{}^{\nu}$. With the known values of these constants the integration of Eqs. (3.2) is straightforward.

The determinant of Eqs. (3.6) becomes singular for $L=0$ and 1. This is shown in Appendix III. To remove the singularity, some of the $C_{kl}{}^{\nu}$ are chosen arbitrarily, and the rest of the $C_{kl}{}^{\nu}$ are found in terms of the chosen ones.

B. Solution at the Origin

In order that the four solutions of **u** be independent of each other we must have

$$\sum_{j=1}^{4}C_j u_{ij}\neq 0, \qquad i=1,2,3,4, \quad (3.8)$$

where C_j are some constants. A necessary condition for this to be satisfied is that the determinant of Eq. (3.8) be nonzero,

$$\|\,u_{ij}\,\|\neq 0. \quad (3.9)$$

It is not difficult to see that this also is a sufficient condition. At the origin the solution u_{ij} can be

expressed as power series in r,

$$u_{ij}=\sum_{\nu=0}^{\infty}a_{ij}{}^{\nu}r^{s_i+\nu}, \tag{3.10}$$

where $a_{ij}{}^{\nu}$ are the coefficients of expansion, and s_i are given integers for each component of **u** and are fixed by the behavior of Eq. (2.23) at the origin. We can satisfy Eq. (3.9) at the origin by having

$$||\, a_{ij}{}^{0}\,|| \neq 0. \tag{3.11}$$

By choosing suitable values of $a_{ij}{}^{0}$, subject to the restriction (3.11), four independent solutions are obtained.

C. Solution at Large r

With given initial values the solution of Eq. (2.23) can be extended from origin to any desired value of r. In order to obtain the asymptotic amplitudes and the phase shifts, the presence of the centrifugal and long range potentials make it necessary to extend the solutions to infinity. This is undesirable because of the time consumption on the computer, and the accumulated error due to the long range integration. Seaton[23] has solved the problem of r^{-2} long range potentials occurring in the off diagonal terms of the potential matrix **V** by diagonalizing the asymptotic form of the differential equations (2.23) and the corresponding S matrix. By an inverse transformation, the elements of the original S matrix are found.

Instead, we develop here a perturbation theory which is based on the method described by Mott and Massey.[28] The error in the resulting solution is inversely proportional to the square of the distance from the origin.

Equation (2.23) for large distances of r can be written

$$[(d^2/dr^2)+k_n{}^2]\mathbf{u}(k_nl_n, r)=2\mathbf{U}\mathbf{u}(k_nl_n, r), \tag{3.12}$$

where **U** is the sum of the centrifugal potential matrix and the asymptotic form of the **V** matrix. The elements of **U** are given in Appendix IV. A component of Eq. (3.12) is of the following form:

$$[(d^2/dr^2)+k^2]u(r)=g(r)$$

$$g(r)\ll k^2u(r), \quad g(r)\to 0 \quad \text{as} \quad r\to\infty \tag{3.13}$$

The perturbation theory is applied between some large distance R and infinity. Suppose u vanishes at R, then we have the following boundary condition

$$u(R)=0. \tag{3.14}$$

If we represent the solution of the homogeneous equation by $y(r)$ at infinity we must have

$$y(r)=a\sin(kr-kR)$$

$$u(r)=(a+\Delta a)\sin(kr-kR+\eta), \tag{3.15}$$

where ΔA and η are generated by $g(r)$. Since $g(r)$ is small, we can write

$$u=y(1+\zeta), \tag{3.16}$$

where ζ is a small function. Substitution of Eq. (3.16) in Eq. (3.13) gives

$$(d/dr)[y^2(d\zeta/dr)]=g(r)y, \tag{3.17}$$

where, upon double integration, we obtain

$$\zeta=\int_R^r\frac{dr}{y^2}\int_R^r g(r')y\,dr'. \tag{3.18}$$

The constants of integrations are fixed by the condition (3.14) and the fact that $u'(R)=y'(R)$.

We now integrate Eq. (3.18) by parts,

$$\zeta=\left[\int_R^r g(r)y\,dr\right]\left[\int_R^r\frac{dr}{y^2}\right]-\int_R^r g(r)y\,dr\int_R^r\frac{dr}{y^2}. \tag{3.19}$$

When the integration with respect to y is carried out, and the result is substituted in Eq. (3.17), we obtain

$$u(r)$$

$$=\sin(kr-kR)\left[a+k^{-1}\int_R^r g(r)\cos(kr-kR)\,dr\right]$$

$$+\cos(kr-kR)\left[-k^{-1}\int_R^r g(r)\sin(kr-kR)\,dr\right]. \tag{3.20}$$

Comparison of the second of Eqs. (3.15) and Eq. (3.20) shows that

$$\Delta a=k^{-1}\int_R^{\infty}g(r)\cos(kr-kR)\,dr$$

$$\eta=-(ak)^{-1}\int_R^{\infty}g(r)\sin(kr-kR)\,dr \tag{3.21}$$

These equations are exact. The functions $g(r)$ in

the four differential equations (3.12) are given by

$$g_i(r) = 2\sum_j U_{ij}u_j. \tag{3.22}$$

To first order this can be written by

$$g_i(r) = 2\sum_j a_j U_{ij} \sin(k_j r - k_j R_j). \tag{3.23}$$

where R_j is the last zero of u_j with positive slope. Substitution of this equation in Eq. (3.21) gives

$$\Delta a_i = -\sum_j \frac{a_j}{k_i}\int^{R_i} \cos(k_i r - k_i R_i)\, U_{ij} \times \sin(k_j r - k_j R_j)\, dr$$

$$\eta_i = \sum_j \frac{a_j}{a_i k_i}\int^{R_i} \sin(k_i r - k_i R_i)\, U_{ij} \times \sin(k_j r - k_j R_j)\, dr. \tag{3.24}$$

Δa_i and η_i can easily be calculated by substituting the values of U_{ij} from Appendix IV, integrating the resulting integrals by parts and retaining the leading terms.

The asymptotic amplitudes and phase shifts are given by

$$a_i(\infty) = a_i(R_i) + \Delta a_i$$

$$\delta_i(\infty) = \delta_i(R_i) + \eta_i + [L - \delta(i, 3) + \delta(i, 4)]\pi/2 \tag{3.25}$$

where $a_i(R_i)$ and $\delta_i(R_i)$ are the amplitudes and total phase shifts calculated at R_i by the machine, and where $\delta(i, 3)$ and $\delta(i, 4)$ are the δ functions.

D. Details of the Numerical Integration

Milne's[29] method with variable mesh size and Simpson's[30] rule were used for the integration of the differential equations and evaluations of the integrals respectively. As the solution advances from the origin, the differential equations become less sensitive to the size of the increment, and the error of integration falls below certain small number ϵ. At each value of r the value of the function is found, first with the given value of the increment, and second with the value of increment divided in half. The error of integration is defined as the difference between these two solutions. When the error becomes small the increment is doubled until a maximum value is reached. At some distance R_1, all the exchange potentials and, similarly, all the direct potentials except those representing optically allowed transitions and the $2p$-$2p$ elastic scattering potential became vanishingly small (see Appendix IV). At this distance the set of differential equations is replaced by the simpler set containing only these potentials. The integration is continued until some distance R_2, where the first order solution of the rest of the range of integration is obtained by the method developed in Sec. IIIC. No attempt was made to solve any set of linear equations or any matrix equations, as these equations are solvable by the computer in their original form.

The values of the constants of the numerical integration are given below. h_i and h_f are the initial and the final increment of integration. In some exceptional cases, different values were used.

h_i	1×10^{-5}
h_f	0.05
ϵ	1×10^{-4}
R_1	30
R_2	200

All quantities are in units of Bohr radius except ϵ which is dimensionless.

IV. RESULTS

The four differential equations listed in Appendix I were integrated by the methods described in Sec. III. Based on Eqs. (2.41), (2.37), and (2.36) we define the three relationships,

$$D_{mn} = \left|\sum_{i=1}^{4} k_i a_{im} a_{in} \sin(\delta_{im} - \delta_{in})\right| \Big/ \sum_{i=1}^{4} k_i a_{im} a_{in} \left|\sin(\delta_{im} - \delta_{in})\right|, \qquad m, n = 1, 2, 3, 4,\ m \neq n, \tag{4.1}$$

$$D_{mn}' = \left|S_{mn} - S_{nm}\right| \Big/ \left(\left|S_{mn}\right| + \left|S_{nm}\right|\right) \qquad m, n = 1, 2, 3, 4,\ m \neq n, \tag{4.2}$$

$$D_m'' = \left|\sum_{n=1}^{4} \left|S_{mn}\right|^2 - 1\right| \Big/ \sum_{n=1}^{4} \left|S_{mn}\right|^2 + 1, \qquad m = 1, 2, 3, 4. \tag{4.3}$$

In an exact solution of the four differential equations the right-hand side of these equations should vanish, they can therefore be used to test the accuracy of the numerical integration. As an illustration, the numerical values of D_{mn}, D_{mn}', and D_m'' for the case of $1s$–$2s$–$2p$ eigenstates coupling, $\beta=+1$, $k_1=2.0$, and $L=3$ will be given below

$$D_{12}=1.4\times10^{-3}, \quad D_{13}=2.6\times10^{-4}, \quad D_{14}=1.3\times10^{-3}$$
$$D_{23}=5.1\times10^{-4}, \quad D_{24}=2.2\times10^{-3}, \quad D_{34}=1.8\times10^{-3}$$
$$D_{12}'=7.6\times10^{-4}, \quad D_{13}'=5.1\times10^{-3}, \quad D_{14}'=5.6\times10^{-3}$$
$$D_{23}'=5.4\times10^{-3}, \quad D_{24}'=5.7\times10^{-3}, \quad D_{34}'=1.3\times10^{-3}$$
$$D_1''=1.8\times10^{-4}, \quad D_2''=7.6\times10^{-5}, \quad D_3''=2.5\times10^{-4}, \quad D_4''=4.4\times10^{-6}.$$

When L is large, particularly at the threshold, the cross sections are small and the values of these cross sections are more sensitive to the errors in the numerical integration. As a result the D values become large. In the tables that follow, except for a few cases, the cross sections that are listed have maximum D values less than 1%. In the exceptional cases by combining different independent sets of solutions and different mesh sizes it has been assured that the cross sections listed are accurate to within a few units in their last significant figures.

The cross sections listed in the tables are ob-

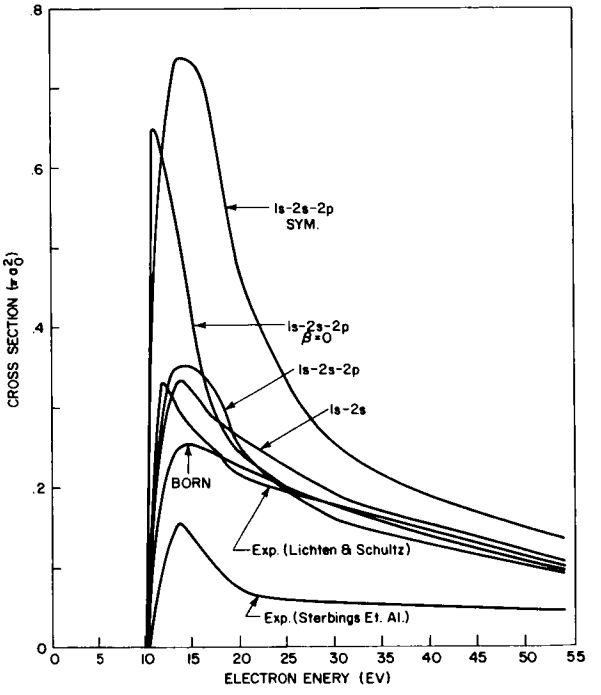


FIGURE 1.—$1s$–$2s$ Excitation cross section. The experimental curves are given in references 1 and 2. The $1s$–$2s$ curve corresponds to a wave function which contains the $1s$ and the $2s$ atomic states. The $1s$–$2s$–$2p$ curve corresponds to the addition of the $2p$ states to the above wave function. The $\beta=0$ curve is the case when the exchange potentials are neglected. In $1s$–$2s$–$2p$ sym it is assumed that electrons are spinless particles whose total wave function is symmetric.

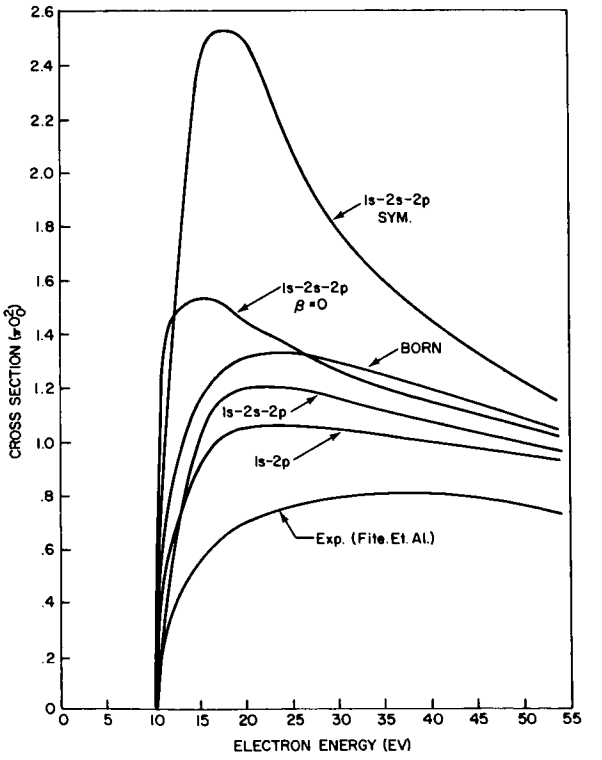


FIGURE 2.—$1s$–$2p$ Excitation cross section. The experimental curve is given in reference 3. The $1s$–$2p$ curve corresponds to a wave function which contains the $1s$ and the $2p$ atomic states. The rest of the curves have the same meaning as in figure 1.

tained by averaging the values of cross sections from two independent sets of solutions of the four differential equations. These independent solutions are obtained by choosing two different values for the determinant (3.11). For small values of k_1 a maximum mesh size of 0.1 in Bohr radius is used while for large values of k_1 this mesh size is 0.05. It is believed that within the framework of the present formalism the cross sections are accurate within one or two units in their last significant figures.

Corresponding to the three channels $1s$, $2s$ and $2p$, the cross-section matrix is a 3×3 matrix. By Eqs. (2.32) and (2.37) the excitation cross sections $2s\rightarrow1s$, $2p\rightarrow1s$ and $2p\rightarrow2s$ can easily be found in terms of the excitation cross sections $1s\rightarrow2s$, $1s\rightarrow2p$ and $2s\rightarrow2p$. The excitation cross section $2s\rightarrow2p$ is quite large and should be calculated by taking into account the energy difference between the levels $2S_{1/2}$, $2P_{1/2}$ and $2P_{3/2}$. The summation over partial cross sections does not converge for this transition and the method of partial wave is not useful. The Born calculation

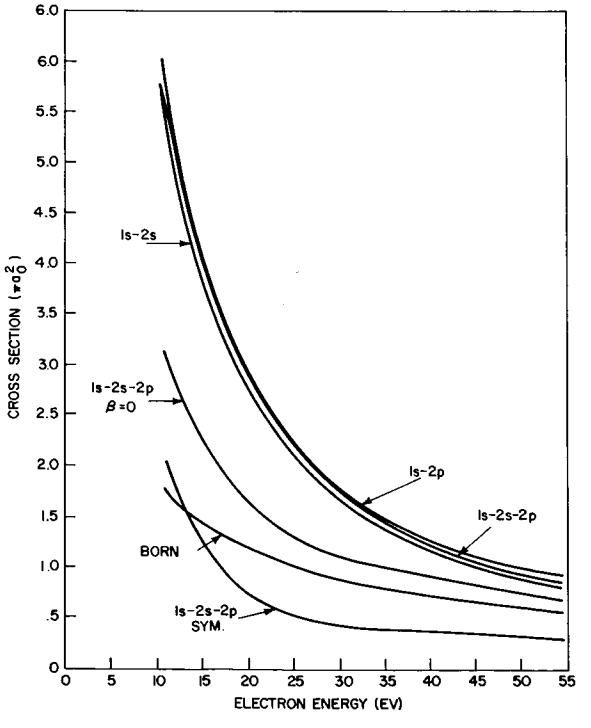


FIGURE 3.—$1s$–$1s$ Elastic cross section. The curves have the same meaning as in figures 1 and 2.

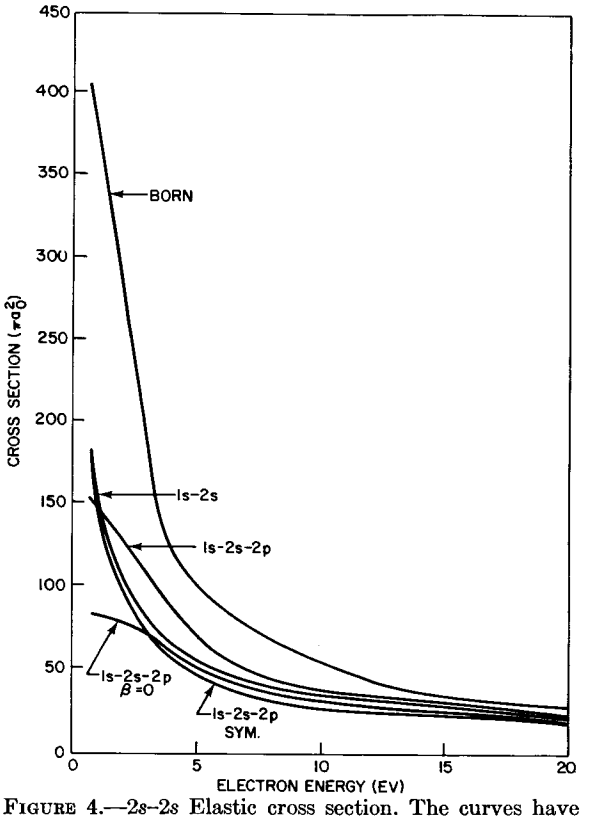


FIGURE 4.—$2s$–$2s$ Elastic cross section. The curves have the same meaning as in figures 1 and 2.

of this cross section is given by Seaton.[31] In the 5 tables that follow the excitation cross sections $1s-2s$, $1s-2p$, and the elastic scattering cross sections $1s-1s$, $2s-2s$ and $2p-2p$ are given. The $2s-2s$, $1s-2p$, and $1s-2s-2p$ eigenstates coupling approximations correspond respectively to the inclusion of the $1s$, $2s$; $1s$, $2p$ and $1s$, $2s$, $2p$ hydrogen eigenstates in the total wave function. The case $\beta=0$ corresponds to the neglect of the exchange terms in the $1s-2s-2p$ eigenstates coupling.

To find the contribution of the higher partial waves whose cross sections are difficult to calculate by the method of eigenstate expansion, we use the regular partial wave Born approximation. Although the partial wave Born approximation as defined by Seaton,[23] in which the continuity of current is satisfied, agrees better with eigenstate expansion calculation, this is not true in general. As a result, for the sake of consistency, we calcu-

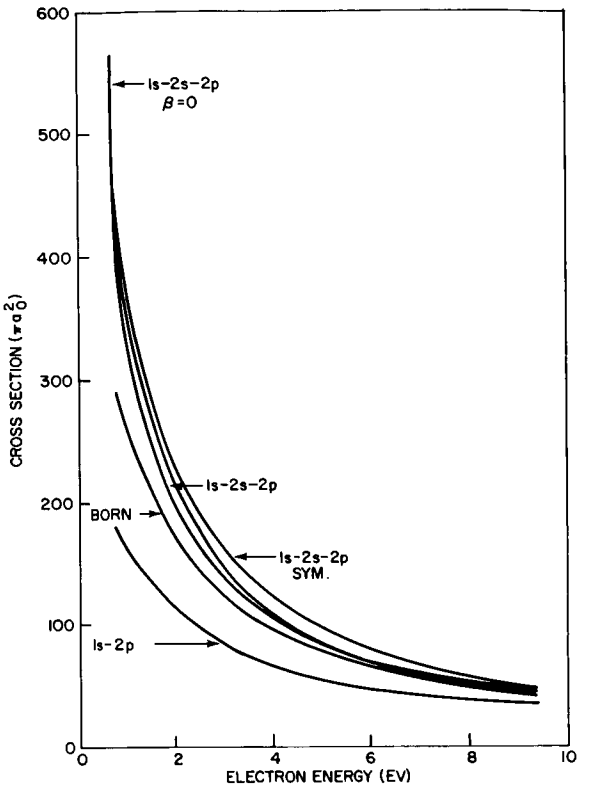


FIGURE 5.—$2p$–$2p$ Elastic cross section. The curves have the same meaning as in figure 2.

lated the contribution of higher partial waves through regular Born approximation.

The cross sections corresponding to the five tables are plotted in the five figures.

V. DISCUSSION

(i) The results of numerical integration using noniterative method agrees with the iterative numerical integration of Burke, Smith and Schey,[7,32] although the two results are different sometimes by as much as 10%. The difference between the two methods is large when the cross section is small.

(ii) $1s$–$2s$ excitation cross section. As is seen from Fig. 1, the $1s$–$2s$ and the $1s$–$2s$–$2p$ eigenstates calculations are in good agreement with the experimental results of Lichten and Schultz[1] although the calculations do not give the sharp peak in the cross section at the threshold as is observed experimentally. The shape of the exchange neglected $1s$–$2s$–$2p$ eigenstates calculation curve is in excellent agreement with the shape of the experimental curve. The two curves coincide beyond 25 ev. We notice that the Born approximation agrees with the eigenstate expansion calculations for values of bombarding energies close to the threshold. The normalization of the experimental results in reference 1 to the Born approximation at 45 ev appears, therefore, to be justified.

(iii) $1s$–$2p$ excitation cross section. The experimental curve[3] agrees in shape with the calculated curves but it is lower than all of them (cf. Fig. 2). It should be noted that the inclusion of the $2s$ state in the $1s$–$2p$ excitation cross section calculation increases this cross section.

Since the experimental curve of Stebbings et al. in Fig. 1 is normalized according to the experimental curve in Fig. 2, if in a new measurement of the $1s$–$2p$ transition cross section higher values for the cross section is found, the two experimental curves in Fig. 1 will be brought into better agreement.

(iv) $2s$–$2s$ elastic cross section. The cross section at the threshold is large. A Ramsauer effect is evident in the $L=0$ case (cf. Table IV).

(v) $2p$–$2p$ elastic cross section. The cross section at the threshold is considerably larger than $2s$–$2s$ elastic cross section. The Born approximation agrees with the eigenstate expansion calculations at higher energies. This cross section is not calculated in reference 32.

(vi) Although we have neglected the spin orbit interaction we have investigated the role that spin plays in electron atom collision. Assuming that electrons are identical but spinless particles, the total wave function should be symmetric with respect to the coordinate interchange of the two electrons. The curves $1s$–$2s$–$2p$ symmetric in the five figures correspond to such a wave function. The disagreement between theory and experiment is an indication of the important role the antisymmetry property of the total electron wave function plays in the free bound electron collisions.

(vii) A study of the five figures shows that the Born approximation, despite its simplicity, agrees with eigenstate expansion calculations for low and close to the thresholds bombarding electron energies. Of particular interest is the fact, yet

unexplained, that the Born approximation underestimates some cross sections and overestimates others.

VI. CONCLUSION

The hydrogen eigenstate expansion of the total wave function of the system of a free electron in collision with a hydrogen atom seems to be the most natural expansion in any partial wave calculation. The formalism can easily be extended to the electron impact atomic hydrogen excitation cross section calculation higher than $n=2$, and to the electron impact excitation cross section of any hydrogenic atom whose central potential field is given by a potential such as Hartree or Hartree–Fock potentials. Where the Born approximation does not give accurate results the eigenstate expansion seems at present to be the most suitable method.

In addition the treatment of the angular momentum as described here is the same for any system of one free and two bound particles with radial interacting forces, and the formalism developed here can be applied to such problems.

We conclude that the Born approximation can be used with more reliability in electron impact atomic excitation cross sections.

Finally, it is believed that the noniterative technique employed here to solve the radial differential equations will find more applications in problems containing exchange integrals.

ACKNOWLEDGMENT

The programming of the numerical integration of the radial differential equations on IBM 7090 computer has been performed by Mr. Edward Sullivan. Through his meticulous and systematic programming the solution of the present problem has become available.

I am indebted to Dr. A. Temkin for many fruitful and illuminating discussions. I should thank Professor Myers of the University of Maryland for a clarifying discussion.

Appendix I

THE FOUR DIFFERENTIAL EQUATIONS

$$\left[\frac{d^2}{dr^2}+k_1{}^2-\frac{L(L+1)}{r^2}+\frac{2}{r}\right]u(k_1L,r)=2y_0(1s1s,r)u(k_1L,r)+2y_0(1s2s,r)u(k_2L,r)$$

$$+2\left[\frac{L}{3(2L+1)}\right]^{1/2}y_1(1s2p,r)u(k_2L-1,r)-2\left[\frac{L+1}{3(2L+1)}\right]^{1/2}y_1(1s2p,r)u(k_2L+1,r)$$

$$+\frac{2\beta}{2L+1}rR_{10}(r)y_L(1sk_1L,r)+\frac{2\beta}{2L+1}rR_{20}(r)y_L(1sk_2L,r)$$

$$+2\beta\left[\frac{3L}{(2L+1)(2L-1)^2}\right]^{1/2}rR_{21}(r)y_{L-1}(1sk_2L-1,r)-2\beta\left[\frac{3(L+1)}{(2L+1)(2L+3)^2}\right]^{1/2}rR_{21}(r)y_{L+1}(1sk_2L+1,r)$$

$$-\beta\delta(L,0)(1+k_1{}^2)(1s\mid k_1L)rR_{10}(r)-\beta\delta(L,0)(1+k_2{}^2)(1s\mid k_2L)rR_{20}(r)$$

$$-\beta\delta(L-1,0)(1+k_2{}^2)(1s\mid k_2L-1)rR_{21}(r). \qquad \text{(A1)}$$

$$\left[\frac{d^2}{dr^2}+k_2{}^2-\frac{L(L+1)}{r^2}+\frac{2}{r}\right]u(k_2L,r)=2y_0(1s2s,r)u(k_1L,r)+2y_0(2s2s,r)u(k_2L,r)$$

$$+2\left[\frac{L}{3(2L+1)}\right]^{1/2}y_1(2s2p,r)u(k_2L-1,r)-2\left[\frac{L+1}{3(2L+1)}\right]^{1/2}y_1(2s2p,r)u(k_2L+1,r)$$

$$+\frac{2\beta}{2L+1}rR_{10}(r)y_L(2sk_1L,r)+\frac{2\beta}{2L+1}rR_{20}(r)y_L(2sk_2L,r)$$

$$+2\beta\left[\frac{3L}{(2L+1)(2L-1)^2}\right]^{1/2}rR_{21}(r)y_{L-1}(2sk_2L-1,r)-2\beta\left[\frac{3(L+1)}{(2L+1)(2L+3)^2}\right]^{1/2}rR_{21}(r)y_{L+1}(2sk_2L+1,r)$$

$$-\beta\delta(L,0)(\tfrac{1}{4}+k_1{}^2)(2s\mid k_1L)rR_{10}(r)-\beta\delta(L,0)(\tfrac{1}{4}+k_2{}^2)(2s\mid k_2L)rR_{20}(r)$$

$$-\beta\delta(L-1,0)(\tfrac{1}{4}+k_2{}^2)(2s\mid k_2L-1)rR_{21}(r). \qquad \text{(A2)}$$

$$\left[\frac{d^2}{dr^2}+k_2^2-\frac{(L-1)L}{r^2}+\frac{2}{r}\right]u(k_2L-1,r)=2\left[\frac{L}{3(2L+1)}\right]^{1/2}y_1(1s2p,r)u(k_1L,r)$$

$$+2\left[\frac{L}{3(2L+1)}\right]^{1/2}y_1(2s2p,r)u(k_2L,r)+2\left[y_0(2p2p,r)+\frac{L-1}{5(2L+1)}y_2(2p2p,r)\right]u(k_2L-1,r)$$

$$-\frac{6}{5}\frac{[L(L+1)]^{1/2}}{2L+1}y_2(2p2p,r)u(k_2L+1,r)+2\beta\left[\frac{3L}{(2L+1)(2L-1)^2}\right]^{1/2}rR_{10}(r)y_{L-1}(2pk_1L,r)$$

$$+2\beta\left[\frac{3L}{(2L+1)(2L-1)^2}\right]^{1/2}rR_{20}(r)y_{L-1}(2pk_2L,r)$$

$$+\frac{6\beta}{2L-1}rR_{21}(r)\left[\frac{y_L(2pk_2L-1,r)}{(2L+1)^2}+\frac{L-1}{2L-3}y_{L-2}(2pk_2L-1,r)\right]-\frac{6\beta[L(L+1)]^{1/2}}{(2L+1)^2}y_L(2pk_2L+1,r)rR_{21}(r)$$

$$-\beta\delta(L,1)\left(\tfrac{1}{4}+k_1^2\right)(2p\mid k_1L)rR_{10}(r)-\beta\delta(L,1)\left(\tfrac{1}{4}+k_2^2\right)(2p\mid k_2L)rR_{20}(r)$$

$$-\beta\delta(L,2)\left(\tfrac{1}{4}+k_2^2\right)(2p\mid k_2L-1)rR_{21}(r). \qquad \text{(A3)}$$

$$\left[\frac{d^2}{dr^2}+k_2^2-\frac{(L+1)(L+2)}{r^2}+\frac{2}{r}\right]u(k_2L+1,r)=-2\left[\frac{L+1}{3(2L+1)}\right]^{1/2}y_1(1s2p,r)u(k_1L,r)$$

$$-2\left[\frac{L+1}{3(2L+1)}\right]^{1/2}y_1(2s2p,r)u(k_2L,r)-\frac{6}{5}\frac{[L(L+1)]^{1/2}}{2L+1}y_2(2p2p,r)u(k_2L-1,r)$$

$$+2\left[y_0(2p2p,r)+\frac{(L+2)y_2(2p2p,r)}{5(2L+1)}\right]u(k_2L+1,r)-2\beta\left[\frac{3(L+1)}{(2L+1)(2L+3)^2}\right]^{1/2}rR_{10}(r)y_{L+1}(2pk_1L,r)$$

$$-2\beta\left[\frac{3(L+1)}{(2L+1)(2L+3)^2}\right]^{1/2}rR_{20}(r)y_{L+1}(2pk_2L,r)-\frac{6\beta[L(L+1)]^{1/2}}{(2L+1)^2}rR_{21}(r)y_L(2pk_2L-1,r)$$

$$+\frac{6\beta}{2L+3}rR_{21}(r)\left[\frac{y_L(2pk_2L+1,r)}{(2L+1)^2}+\frac{(L+2)y_{L+2}(2pk_2L+1,r)}{(2L+5)}\right]$$

$$-\beta\delta(L,0)\left(\tfrac{1}{4}+k_2^2\right)(2p\mid k_2L+1)rR_{21}(r). \qquad \text{(A4)}$$

Appendix II

ELEMENTS OF THE POTENTIAL MATRIX

Elements of D_{ij}:

$$D_{11} = -(1+r^{-1})e^{-2r}, \qquad D_{22} = -\left(r^{-1}+\frac{3}{4}+\frac{r}{4}+\frac{r^2}{8}\right)e^{-r}$$

$$D_{33} = -\left[r^{-1}+\frac{3}{4}+\frac{r}{4}+\frac{r^2}{24}\right]e^{-r}+\frac{6(L-1)}{2L+1}\left[\frac{1}{r^3}-\left(\frac{1}{r^3}+\frac{1}{r^2}+\frac{1}{2r}+\frac{1}{6}+\frac{r}{24}+\frac{r^2}{144}\right)e^{-r}\right]$$

$$D_{44} = -\left[r^{-1}+\frac{3}{4}+\frac{r}{4}+\frac{r^2}{24}\right]e^{-r}+\frac{6(L+2)}{2L+1}\left[\frac{1}{r^3}-\left(\frac{1}{r^3}+\frac{1}{r^2}+\frac{1}{2r}+\frac{1}{6}+\frac{r}{24}+\frac{r^2}{144}\right)e^{-r}\right]$$

$$D_{12} = D_{21} = \frac{2\sqrt{2}}{9}(r+\tfrac{2}{3})\exp(-\tfrac{3}{2}r)$$

$$D_{13} = D_{31} = \frac{128\sqrt{2}}{243}\times\left(\frac{L}{2L+1}\right)^{1/2}\left[\frac{1}{r^2}-\left(\frac{1}{r^2}+\frac{3}{2r}+\frac{9}{8}+\frac{27r}{64}\right)\exp(-\tfrac{3}{2}r)\right]$$

$$D_{14} = D_{41} = -\frac{128\sqrt{2}}{243}\left(\frac{L+1}{2L+1}\right)^{1/2}\left[\frac{1}{r^2}-\left(\frac{1}{r^2}+\frac{3}{2r}+\frac{9}{8}+\frac{27r}{64}\right)\exp(-\tfrac{3}{2}r)\right]$$

$$D_{23} = D_{32} = -3\left(\frac{L}{2L+1}\right)^{1/2}\left[\frac{1}{r^2}-\left(\frac{1}{r^2}+r^{-1}+\frac{1}{2}+\frac{r}{6}+\frac{r^2}{24}\right)e^{-r}\right]$$

$$D_{24} = D_{42} = 3\left(\frac{L+1}{2L+1}\right)^{1/2}\left[\frac{1}{r^2}-\left(\frac{1}{r^2}+r^{-1}+\frac{1}{2}+\frac{r}{6}+\frac{r^2}{24}\right)e^{-r}\right]$$

$$D_{34} = D_{43} = -18\left[\frac{L(L+1)}{(2L+1)^2}\right]^{1/2}\left[\frac{1}{r^3}-\left(\frac{1}{r^3}+\frac{1}{r^2}+\frac{1}{2r}+\frac{1}{6}+\frac{r}{24}+\frac{r^2}{144}\right)e^{-r}\right]$$

Elements of F_{ij}:

$$F_{11} = \frac{\beta}{2L+1}\left[\frac{R_{10}}{r^L}\int_0^r R_{10}r'^{L+1}\,dr' - R_{10}r^{L+1}\int_0^r \frac{R_{10}}{r'^L}\,dr'\right]$$

$$F_{22}=\frac{\beta}{2L+1}\left[\frac{R_{20}}{r^L}\int_0^r R_{20}r'^{L+1}\,dr'-R_{20}r^{L+1}\int_0^r\frac{R_{20}}{r'^L}\,dr'\right]$$

$$F_{33}=\frac{3\beta}{2L-1}\left[[(2L+1)^2]^{-1}\left(\frac{R_{21}}{r^L}\int_0^r R_{21}r'^{L+1}\,dr'-R_{21}r^{L+1}\int_0^r\frac{R_{21}}{r'^L}\,dr'\right)\right.$$

$$\left.+\frac{L-1}{2L-3}\left(\frac{R_{21}}{r^{L-2}}\int_0^r R_{21}r'^{L-1}\,dr'-R_{21}r^{L-1}\int_0^r\frac{R_{21}}{r'^{L-2}}\,dr'\right)\right]$$

$$F_{44}=\frac{3\beta}{2L+3}\left[[(2L+1)^2]^{-1}\left(\frac{R_{21}}{r^L}\int_0^r R_{21}r'^{L+1}\,dr'-R_{21}r^{L+1}\int_0^r\frac{R_{21}}{r'^L}\,dr'\right)\right.$$

$$\left.+\frac{L+2}{2L+5}\left(\frac{R_{21}}{r^{L+2}}\int_0^r R_{21}r'^{L+3}\,dr'-R_{21}r^{L+3}\int_0^r\frac{R_{21}}{r'^{L+2}}\,dr'\right)\right]$$

$$F_{12}=\frac{\beta}{2L+1}\left[\frac{R_{20}}{r^L}\int_0^r R_{10}r'^{L+1}\,dr'-R_{20}r^{L+1}\int_0^r\frac{R_{10}}{r'^L}\,dr'\right]$$

$$F_{21}=F_{12}[R_{10}\rightleftharpoons R_{20}]$$

$$F_{13}=\sqrt{3}\beta\left[\frac{L}{(2L+1)(2L-1)^2}\right]^{1/2}\left[\frac{R_{21}}{r^{L-1}}\int_0^r R_{10}r'^L\,dr-R_{21}r^L\int_0^r\frac{R_{10}}{r'^{L-1}}\,dr'\right]$$

$$F_{31}=F_{13}[R_{10}\rightleftharpoons R_{21}]$$

$$F_{14}=-\sqrt{3}\beta\left[\frac{L+1}{(2L+1)(2L+3)^2}\right]^{1/2}\left[\frac{R_{21}}{r^{L+1}}\int_0^r R_{10}r'^{L+2}\,dr'-R_{21}r^{L+2}\int_0^r\frac{R_{10}}{r'^{L+1}}\,dr'\right]$$

$$F_{41}=F_{14}[R_{10}\rightleftharpoons R_{21}]$$

$$F_{23}=\sqrt{3}\beta\left[\frac{L}{(2L+1)(2L-1)^2}\right]^{1/2}\left[\frac{R_{21}}{r^{L-1}}\int_0^r R_{20}r'^L\,dr'-R_{21}r^L\int_0^r\frac{R_{20}}{r'^{L-1}}\,dr'\right]$$

$$F_{32}=F_{23}[R_{20}\rightleftharpoons R_{21}]$$

$$F_{24}=-\sqrt{3}\beta\left[\frac{L+1}{(2L+1)(2L+3)^2}\right]^{1/2}\left[\frac{R_{21}}{r^{L+1}}\int_0^r R_{20}r'^{L+2}\,dr'-R_{21}r^{L+2}\int_0^r\frac{R_{20}}{r'^{L+1}}\,dr'\right]$$

$$F_{42}=F_{24}[R_{20}\rightleftharpoons R_{21}]$$

$$F_{34}=-3\beta\left[\frac{L(L+1)}{(2L+1)^4}\right]^{1/2}\left[\frac{R_{21}}{r^L}\int_0^r R_{21}r'^{L+1}\,dr'-R_{21}r^{L+1}\int_0^r\frac{R_{21}}{r'^L}\,dr'\right]$$

$$F_{34}=F_{43}$$

Elements of g_{ij} and h_{ij}:

$$g_{11}=\beta R_{10}r^{L+1}/2L+1, \qquad h_{11}=R_{10}[(r^L)^{-1}-\tfrac{1}{2}(1+k_1^2)\,\delta(L,0)r]$$

$$g_{22}=\beta R_{20}r^{L+1}/2L+1, \qquad h_{22}=R_{20}[(r^L)^{-1}-\tfrac{1}{2}(\tfrac{1}{4}+k_2^2)\,\delta(L,0)r]$$

$$g_{33}{}^1=\frac{3\beta R_{21}r^{L+1}}{(2L-1)\,(2L+1)^2}, \qquad h_{33}{}^1=\frac{R_{21}}{r^L}$$

$$g_{33}{}^2=\frac{3\beta(L-1)\,R_{21}r^{L-1}}{(2L-1)\,(2L-3)}, \qquad h_{33}{}^2=R_{21}[(r^{L-2})^{-1}-\tfrac{1}{2}(\tfrac{1}{4}+k_2^2)\,\delta(L,2)r]$$

$$g_{44}{}^1=\frac{3\beta R_{21}r^{L+1}}{(2L+3)\,(2L+1)^2}, \qquad h_{44}{}^1=R_{21}[(r^L)^{-1}-\tfrac{1}{2}(\tfrac{1}{4}+k_2^2)\,\delta(L,0)r]$$

$$g_{44}{}^2=\frac{3\beta(L+2)\,R_{21}r^{L+3}}{(2L+3)\,(2L+5)}, \qquad h_{44}{}^2=R_{21}/r^{L+2}$$

$$g_{12}=[\beta/(2L+1)]R_{20}r^{L+1}, \qquad h_{12}=R_{10}[(r^L)^{-1}-\tfrac{1}{2}(1+k_2^2)\,\delta(L,0)r]$$

$$g_{21}=g_{12}[R_{20}\rightarrow R_{10}], \qquad h_{21}=k_{12}[R_{10}\rightarrow R_{20}]$$

$$g_{13}=\sqrt{3}\beta\left[\frac{L}{(2L+1)\,(2L-1)^2}\right]^{1/2}R_{21}r^L, \qquad h_{13}=R_{10}[(r^{L-1})^{-1}-\tfrac{1}{2}(1+k_2^2)\,\delta(L,1)r]$$

$$g_{31}=g_{13}[R_{21}\rightarrow R_{10}], \qquad h_{31}=h_{13}[R_{10}\rightarrow R_{21}]$$

$$g_{14}=-\sqrt{3}\beta\left[\frac{L+1}{(2L+1)\,(2L+3)^2}\right]^{1/2}R_{21}r^{L+2}, \qquad h_{14}=R_{10}/r^{L+1}$$

$$g_{41}=g_{14}[R_{21}\rightarrow R_{10}], \qquad h_{41}=h_{14}[R_{10}\rightarrow R_{21}]$$

$$g_{23}=\sqrt{3}\beta\left[\frac{L}{(2L+1)\,(2L-1)^2}\right]^{1/2}R_{21}r^L, \qquad h_{23}=R_{20}[(r^{L-1})^{-1}-\tfrac{1}{2}(\tfrac{1}{4}+k_2^2)\,\delta(L,1)r]$$

$$g_{32}=g_{23}[R_{21}\rightarrow R_{20}], \qquad h_{32}=h_{23}[R_{20}\rightarrow R_{21}]$$

$$g_{24}=-\sqrt{3}\beta\left[\frac{L+1}{(2L+1)\,(2L+3)^2}\right]^{1/2}R_{21}r^{L+2}, \qquad h_{24}=R_{20}/r^{L+1}$$

$$g_{42}=g_{24}[R_{21}\rightarrow R_{20}], \qquad h_{42}=h_{24}[R_{20}\rightarrow R_{21}]$$

$$g_{34} = -3\beta\left[\frac{L(L+1)}{(2L+1)^4}\right]^{1/2} R_{21} r^{L+1}, \qquad h_{34} = R_{21}/r^L$$

$$g_{43} = g_{34}$$

In F_{ij} matrix the interchange of the functions R_{10}, R_{20} and R_{21} accompanies the interchange of their arguments, too.

Appendix III

SINGULARITY OF THE DETERMINANT OF EQUATION (3.6) FOR $L=0$ AND 1

(1) $L=0$ Case. By making use of the definition of D_{ij} and F_{ij} and Eq. (3.7), the following relationship can be derived from Eq. (3.3)

$$\int_0^\infty \{rR_{20}[(d^2/dr^2)+k_1^2]v_1-\beta rR_{10}[(d^2/dr^2)+k_2^2]v_2\}\,dr=-(2/\sqrt{3})[a_{13}B_{24}-\beta a_{23}B_{14}], \tag{A5}$$

where

$$a_{13}=\int_0^\infty R_{10}R_{21}r^3\,dr=[2^{15}\times 3^{-9}]^{1/2}$$

$$a_{23}=\int_0^\infty R_{20}R_{21}r^3\,dr=-3\sqrt{3}$$

Integrating the left-hand side of Eq. (A5) by parts, and making use of Eqs. (2.11) and (3.7) we obtain

$$\int_0^\infty \{rR_{20}[(d^2/dr^2)+k_1^2]v_1-\beta rR_{10}[(d^2/dr^2)+k_2^2]v_2\}\,dr=-2[B_{21}-\beta B_{12}]. \tag{A6}$$

We conclude

$$B_{21}-\beta B_{12}=(1/\sqrt{3})[a_{13}B_{24}-\beta a_{23}B_{14}]. \tag{A7}$$

Therefore, all the elements on the right-hand side of Eqs. (3.6) are not independent. In order that these equations be consistent, one of them should be a linear combination of others. By making use of the first of Eqs. (3.7), it can be shown directly that equations similar to Eq. (A7) hold among the elements of each row of the determinant of Eqs. (3.6).

(11) $L=1$ Case. Similar to the previous case, the following relationship can be derived from Eqs. (3.3)

$$\int_0^\infty \{rR_{21}[(d^2/dr^2)+k_1^2-(2/r^2)]v_1-\beta rR_{10}[(d^2/dr^2)+k_2^2]v_3\}\,dr$$

$$=-\tfrac{2}{3}[\beta a_{13}B_{11}+\beta a_{23}B_{12}-a_{13}B_{33}{}^1+\sqrt{2}(a_{13}B_{34}-\tfrac{3}{5}\beta a_{33}B_{14})], \tag{A8}$$

where

$$a_{33}=\int_0^\infty R_{21}{}^2r^4\,dr=30.$$

Integrating the left-hand side of Eq. (A8) by parts, and making use of Eqs. (2.11) and (3.7) we obtain

$$\int_0^\infty \{rR_{21}[(d^2/dr^2)+k_1^2-(2/r^2)]v_1-\beta rR_{10}[(d^2/dr^2)+k_2^2]v_3\}\,dr=-2[B_{31}-\beta B_{13}]. \tag{A9}$$

Combining Eqs. (A8) and (A9) we get

$$B_{31}-\beta B_{13}=\tfrac{1}{3}[\beta a_{13}B_{11}+\beta a_{23}B_{12}-a_{13}B_{33}{}^1+\sqrt{2}(a_{13}B_{34}-\tfrac{3}{5}\beta a_{33}B_{14})] \tag{A10}$$

Finally, Eqs. (3.3) give the following relationship

$$\int_0^\infty \{rR_{21}[(d^2/dr^2)+k_2^2-(2/r^2)]v_2-\beta rR_{20}[(d^2/dr^2)+k_2^2]v_3\}\,dr$$

$$=-\tfrac{2}{3}[\beta a_{23}B_{22}+\beta a_{13}B_{21}-a_{23}B_{33}{}^1+\sqrt{2}(a_{23}B_{34}-\tfrac{3}{5}\beta a_{33}B_{24})] \tag{A11}$$

Integration by parts of the left-hand side, as before, gives

$$\int_0^\infty \{rR_{21}[(d^2/dr^2)+k_2^2-(2/r^2)]v_2-\beta rR_{20}[(d^2/dr^2)+k_2^2]v_3\}\,dr=-2[B_{32}-\beta B_{23}], \tag{A12}$$

whereupon we get

$$B_{32}-\beta B_{23}=\tfrac{1}{3}[\beta a_{23}B_{22}+\beta a_{13}B_{21}-a_{23}B_{33}{}^1+\sqrt{2}(a_{23}B_{34}-\tfrac{3}{5}\beta a_{33}B_{24})] \tag{A13}$$

To remove the singularity in $L=0$ case one of the $C_{kl}{}'$ is chosen arbitrary, and a degenerate equation is removed from Eqs. (3.6). Similarly in $L=1$ case, two of the $C_{kl}{}'$ are chosen arbitrary and two degenerate equations are removed from Eqs. (3.6).

Appendix IV

ELEMENTS OF THE MATRIX OF THE SUM OF THE ASYMPTOTIC COULOMB AND CENTRIFUGAL POTENTIALS

$U_{11}=L(L+1)r^{-2}$,

$U_{22}=L(L+1)r^{-2}$

$U_{33}=(L-1)Lr^{-2}+12(L-1)(2L+1)^{-1}r^{-3}$,

$U_{44}=(L+1)(L+2)r^{-2}+12(L+2)(2L+1)^{-1}r^{-3}$

$U_{12}=U_{21}=0$,

$U_{13}=U_{31}=[256\sqrt{2}/243][L/(2L+1)]^{1/2}r^{-2}$

$U_{14}=U_{41}=-[256\sqrt{2}/243][(L+1)/(2L+1)]^{1/2}r^{-2}$,

$U_{23}=U_{32}=-6[L/(2L+1)]^{1/2}r^{-2}$

$U_{24}=U_{42}=6[(L+1)/(2L+1)]^{1/2}r^{-2}$,

$U_{34}=U_{43}=-36[L(L+1)]^{1/2}(2L+1)^{-1}r^{-3}$

TABLES OF CROSS SECTIONS

In the following five tables the partial and the total cross sections for the processes $1s$–$2s$, $1s$–$2p$, $1s$–$1s$, $2s$–$2s$ and $2p$–$2p$ in different approximations are listed. Each partial cross section corresponds to a given total orbital angular momentum L which appears at the head of each column. $\sum_S$ is the sum of the singlet and $\sum_T$ is the sum of the triplet partial cross sections within certain approximation and for a given electron energy. Q_T is the total cross section, and in the case of the eigenstates coupling, Born approximation is used to add the contribution of the partial waves beyond the maximum L which has been computed within the given eigenstates coupling approximation. This contribution can easily be obtained by using the given table of the Born approximation of the partial and the total cross sections. k_1 is the wave number of the electron beam incident on the ground state of the atom and k_2 is the wave number of the same beam when the atom is in its first excited state. The energy of the beam is given in electron volt through

$$E = 13.6k^2,$$

where k can be k_1 or k_2. In Table V, l_1 and l_2 are the quantum numbers of the orbital angular momenta of the bound and the free electrons respectively.

TABLE I.—1*s*–2*s* *Excitation Cross Sections*

A. Born approximation

k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum$	Q_T
0.9	0.16376	0.00981	0.00021	0.00000	0.00000	0.00000	0.00000	0.00000	0.17378	0.17379
1.0	0.19578	0.04795	0.00428	0.00026	0.00000	0.00000	0.00000	0.00000	0.24827	0.24827
1.1	0.16272	0.07073	0.01141	0.00125	0.00011	0.00001	0.00000	0.00000	0.24622	0.24623
1.2	0.12704	0.07896	0.01858	0.00299	0.00039	0.00004	0.00000	0.00000	0.22800	0.22800
1.5	0.05872	0.06606	0.02979	0.00939	0.00242	0.00053	0.00010	0.00002	0.16703	0.16706
2.0	0.01946	0.03363	0.02521	0.01365	0.00614	0.00236	0.00081	0.00025	0.10151	0.10187
3.0	0.00388	0.00909	0.01019	0.00866	0.00628	0.00394	0.00226	0.00116	0.04546	0.04758
4.0	0.00123	0.00320	0.00420	0.00431	0.00385	0.00301	0.00216	0.00141	0.02337	0.02720

B. Exchange neglected 1*s*–2*s*–2*p* eigenstates coupling approximation

k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum$	Q_T
0.9	0.2202	0.0749	0.3535						0.6486	0.6486
1.0	0.1685	0.1427	0.1598	0.0507					0.5227	0.5227
1.1	0.0951	0.1142	0.0298	0.0616	0.0231				0.3238	0.3238
1.2	0.0594	0.1137	0.0032	0.0360	0.0244	0.0135			0.2502	0.2502
1.5	0.0249	0.0861	0.0201	0.0068	0.0112	0.0118	0.0074		0.1683	0.1683
2.0	0.0101	0.0373	0.0255	0.0107	0.0046	0.0034	0.0033		0.0949	0.0953

C. 1*s*–2*p* eigenstates coupling approximation

Singlet

k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_S$
0.9	0.0375	0.0017	0.0000	0.0000	0.0000	0.0000	0.0000		0.0392
1.0	0.0725	0.0583	0.0002	0.0000	0.0000	0.0000	0.0000		0.1310
1.1	0.0701	0.0525	0.0023	0.0000	0.0000	0.0000	0.0000		0.1249
1.2	0.0547	0.0534	0.0054	0.0002	0.0000	0.0000	0.0000		0.1137
1.5	0.0241	0.0384	0.0110	0.0022	0.0004	0.0001	0.0000		0.0762
2.0	0.0072	0.0157	0.0093	0.0041	0.0015	0.0005	0.0002		0.0385

Triplet

k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_T$	$\sum_S+\sum_T$	Q_T
0.9	0.0004	0.1686	0.0060	0.0000	0.0000	0.0000	0.0000		0.1750	0.2142	0.2142
1.0	0.0021	0.1528	0.0446	0.0021	0.0001	0.0000	0.0000		0.2017	0.3327	0.3327
1.1	0.0044	0.1052	0.0568	0.0068	0.0005	0.0000	0.0000		0.1737	0.2987	0.2987
1.2	0.0061	0.0737	0.0576	0.0114	0.0015	0.0002	0.0000		0.1505	0.2642	0.2642
1.5	0.0073	0.0355	0.0406	0.0174	0.0050	0.0012	0.0002		0.1072	0.1833	0.1833
2.0	0.0049	0.0162	0.0205	0.0143	0.0074	0.0032	0.0012		0.0677	0.1062	0.1068

TABLE I.—Continued

D. 1s–2s–2p eigenstates coupling approximation

	Singlet										
k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_S$		
0.9	0.0523	0.0048	0.0620						0.1191		
1.0	0.0768	0.0147	0.0833	0.0092					0.1840		
1.1	0.0585	0.0245	0.0647	0.0236	0.0055				0.1768		
1.2	0.0382	0.0251	0.0246	0.0252	0.0081	0.0028			0.1240		
1.5	0.0123	0.0308	0.0015	0.0041	0.0051	0.0034	0.0023	0.0026	0.0621		
2.0	0.0049	0.0152	0.0068	0.0021	0.0010	0.0010	0.0011	0.0008	0.0329		
3.0	0.0010	0.0031	0.0031	0.0023	0.0015	0.0009	0.0005	0.0006	0.0130		
4.0	0.0003	0.0010	0.0012	0.0012	0.0010	0.0008			0.0055		
	Triplet										
k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_T$	$\sum_S+\sum_T$	Q_T
0.9	0.0013	0.0748	0.0019						0.0780	0.1971	0.1971
1.0	0.0040	0.1224	0.0195	0.0214					0.1673	0.3513	0.3513
1.1	0.0050	0.1013	0.0326	0.0077	0.0131				0.1597	0.3366	0.3366
1.2	0.0055	0.0724	0.0359	0.0036	0.0105	0.0076			0.1355	0.2596	0.2596
1.5	0.0045	0.0333	0.0309	0.0072	0.0046	0.0054	0.0049	0.0043	0.0951	0.1573	0.1573
2.0	0.0031	0.0155	0.0176	0.0101	0.0044	0.0025	0.0023	0.0019	0.0574	0.0903	0.0907
3.0	0.0013	0.0048	0.0065	0.0059	0.0044	0.0029	0.0018	0.0012	0.0288	0.0418	0.0439
4.0	0.0006	0.0019	0.0028	0.0030	0.0027	0.0022			0.0132	0.0187	0.0261

TABLE II.—1s–2p *Excitation Cross Sections*

A. Born approximation

k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum$	Q_T
0.9	0.00107	0.46700	0.09607	0.01025	0.00087	0.00007	0.00001	0.00000	0.56534	0.57535
1.0	0.00499	0.48867	0.35645	0.13467	0.03964	0.01032	0.00254	0.00059	1.03787	1.03851
1.1	0.00702	0.36207	0.41698	0.24992	0.11533	0.04649	0.01753	0.00624	1.22158	1.22859
1.2	0.00747	0.25540	0.38137	0.29903	0.17908	0.09269	0.04453	0.02004	1.27961	1.30741
1.5	0.00550	0.09184	0.19986	0.23775	0.21386	0.16292	0.11291	0.07232	1.09696	1.28101
2.0	0.00234	0.02222	0.05938	0.09386	0.11287	0.11376	0.10248	0.08374	0.59065	1.04055
3.0	0.00048	0.00285	0.00806	0.01537	0.02309	0.02920	0.03271	0.03273	0.14449	0.66256
4.0	0.00013	0.00066	0.00179	0.00359	0.00586	0.00816	0.01009	0.01114	0.04142	0.45252

B. Exchange neglected 1s–2s–2p eigenstates coupling approximation

k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum$	Q_T
0.9	0.1600	0.3985	0.6497						1.2082	1.2194
1.0	0.1007	0.2917	0.8190	0.2190	0.0476	0.0224			1.5004	1.5041
1.1	0.0980	0.2008	0.6201	0.3696	0.1380	0.0586	0.0166		1.5017	1.5150
1.2	0.0822	0.1251	0.4481	0.3922	0.2044	0.1014	0.0403		1.3937	1.4416
1.5	0.0372	0.0334	0.1671	0.2568	0.2295	0.1706	0.1136		1.0082	1.2645
2.0	0.0105	0.0068	0.0394	0.0837	0.1089	0.1165	0.1042		0.4700	1.0036

C. 1s–2s eigenstates coupling approximation

	Singlet									
k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_S$	
0.9	0.0044	0.1216	0.1422						0.2682	
1.0	0.0168	0.0655	0.3011	0.0206	0.0057	0.0059			0.4156	
1.1	0.0299	0.0366	0.3948	0.0851	0.0260	0.0099	0.0060		0.5883	
1.2	0.0296	0.0169	0.3088	0.1421	0.0517	0.0236	0.0103		0.5330	
1.5	0.0059	0.0037	0.0821	0.0989	0.0718	0.0458	0.0304		0.3386	
2.0	0.0010	0.0006	0.0131	0.0271	0.0327	0.0335	0.0296		0.1376	

	Triplet										
k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_T$	$\sum_S+\sum_T$	Q_T
0.9	0.0002	0.2066	0.0005						0.2073	0.4755	0.4867
1.0	0.0016	0.1078	0.0020	0.1651	0.0363	0.0187			0.3315	0.7471	0.7508
1.1	0.0037	0.0540	0.0060	0.1599	0.1002	0.0365	0.0217		0.3820	0.9703	0.9836
1.2	0.0055	0.0249	0.0098	0.1446	0.1336	0.0695	0.0341		0.4220	1.0050	1.0529
1.5	0.0059	0.0027	0.0133	0.0868	0.1231	0.1100	0.0849		0.4267	0.7652	1.0215
2.0	0.0028	0.0002	0.0073	0.0311	0.0559	0.0700	0.0772		0.2445	0.3820	0.9156

TABLE II.—Continued

D. 1s–2p eigenstates coupling approximation

	Singlet										
k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_S$		
0.9	0.0390	0.0745	0.1027						0.2162		
1.0	0.0360	0.1123	0.2575	0.0317					0.4375		
1.1	0.0358	0.1094	0.3405	0.0886	0.0308	0.0113	0.0075		0.6239		
1.2	0.0345	0.0806	0.2912	0.1278	0.0506	0.0237	0.0105		0.6189		
1.5	0.0172	0.0175	0.0953	0.1003	0.0693	0.0440	0.0290	0.0238	0.3964		
2.0	0.0036	0.0023	0.0170	0.0303	0.0344	0.0333	0.0293	0.0229	0.1731		
3.0	0.0004	0.0002	0.0015	0.0037	0.0060	0.0078	0.0094	0.0106	0.0396		
4.0	0.0001	0.0001	0.0003	0.0008	0.0014	0.0020			0.0047		
	Triplet										
k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_T$	$\sum_S+\sum_T$	Q_T
0.9	0.0007	0.0682	0.0112						0.0801	0.2963	0.3075
1.0	0.0033	0.0801	0.0500	0.1730					0.3064	0.7439	0.7976
1.1	0.0070	0.0626	0.0567	0.1841	0.1082	0.0404	0.0209		0.4799	1.1039	1.1172
1.2	0.0096	0.0418	0.0537	0.1761	0.1409	0.0729	0.0332		0.5282	1.1471	1.1950
1.5	0.0107	0.0131	0.0351	0.1081	0.1343	0.1157	0.0842	0.0596	0.5608	0.9570	1.1410
2.0	0.0053	0.0038	0.0143	0.0393	0.0625	0.0731	0.0738	0.0642	0.3363	0.5094	0.9593
3.0	0.0010	0.0006	0.0025	0.0066	0.0122	0.0175	0.0220	0.0246	0.0870	0.1266	0.6095
4.0	0.0002	0.0002	0.0006	0.0016	0.0031	0.0048			0.0105	0.0152	0.4475

TABLE III.—1s–1s *Elastic Cross Sections*

A. Born approximation

k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum$	Q_T
0.9	1.65113	0.09722	0.00379	0.00013	0.00000	0.00000	0.00000	0.00000	1.75227	1.75233
1.0	1.42354	0.11192	0.00587	0.00027	0.00001	0.00000	0.00000	0.00000	1.54161	1.54167
1.1	1.22728	0.12348	0.00835	0.00049	0.00003	0.00000	0.00000	0.00000	1.35963	1.35968
1.2	1.05936	0.13181	0.01109	0.00082	0.00006	0.00000	0.00000	0.00000	1.20314	1.20318
1.5	0.69140	0.14017	0.01955	0.00241	0.00028	0.00003	0.00000	0.00000	0.85384	0.85389
2.0	0.36283	0.12285	0.02924	0.00621	0.00124	0.00023	0.00004	0.00001	0.52265	0.52267
3.0	0.12662	0.07286	0.03013	0.01131	0.00402	0.00133	0.00042	0.00012	0.24681	0.24700
4.0	0.05565	0.04195	0.02291	0.01145	0.00546	0.00243	0.00104	0.00042	0.14131	0.14194

B. Exchange neglected 1s–2s–2p eigenstates coupling approximation

k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum$	Q_T
0.9	2.632	0.3300	0.1916	0.0157					3.169	3.169
1.0	2.091	0.2518	0.0890	0.0309	0.0078	0.0028			2.473	2.473
1.1	1.694	0.2302	0.0483	0.0299	0.0105	0.0038	0.0016		2.018	2.018
1.2	1.390	0.2228	0.0346	0.0248	0.0113	0.0045	0.0019		1.690	1.690
1.5	0.8181	0.2032	0.0310	0.0135	0.0091	0.0048	0.0027		1.082	1.082
2.0	0.4053	0.1742	0.0494	0.0181	0.0108	0.0036	0.0027		0.6641	0.6641

C. 1s–2s eigenstates coupling approximation

Singlet

k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_S$
0.9	0.4324	0.0382	0.0015	0.0000	0.0000	0.0000	0.0000		0.4721
1.0	0.2824	0.0338	0.0014	0.0001	0.0000	0.0000	0.0000		0.3177
1.1	0.1865	0.0199	0.0011	0.0001	0.0000	0.0000	0.0000		0.2076
1.2	0.1397	0.0111	0.0006	0.0001	0.0000	0.0000	0.0000		0.1515
1.5	0.0905	0.0012	0.0001	0.0000	0.0000	0.0000	0.0000		0.0918
2.0	0.0608	0.0083	0.0016	0.0003	0.0001	0.0000	0.0000		0.0711

Triplet

k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_T$	$\sum_S+\sum_T$	Q_T
0.9	3.684	1.394	0.0477	0.0016	0.0001	0.0000	0.0000		5.127	5.599	5.599
1.0	2.903	1.162	0.0579	0.0023	0.0001	0.0000	0.0001		4.125	4.443	4.443
1.1	2.297	0.9654	0.0651	0.0036	0.0002	0.0000	0.0000		3.331	3.539	3.539
1.2	1.829	0.8085	0.0696	0.0050	0.0004	0.0000	0.0000		2.713	2.864	2.864
1.5	0.9716	0.4857	0.0727	0.0092	0.0011	0.0001	0.0000		1.5404	1.632	1.632
2.0	0.4042	0.2366	0.0627	0.0142	0.0031	0.0006	0.0001		0.7215	0.7926	0.7926

TABLE III.—Continued

D. $1s$–$2s$–$2p$ eigenstates coupling approximation

	Singlet										
k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_S$		
0.9	0.4230	0.0186	0.0442	0.0043					0.4901		
1.0	0.2984	0.0167	0.0635	0.0073	0.0017	0.0006	0.0005		0.3887		
1.1	0.2148	0.0108	0.0578	0.0104	0.0026	0.0008	0.0003		0.2975		
1.2	0.1569	0.0048	0.0338	0.0105	0.0030	0.0011	0.0005		0.2106		
1.5	0.1040	0.0008	0.0035	0.0038	0.0025	0.0012	0.0007		0.1165		
2.0	0.0698	0.0138	0.0030	0.0018	0.0018	0.0008	0.0006		0.0916		
	Triplet										
k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_T$	$\sum_S+\sum_T$	Q_T
0.9	3.689	1.421	0.0682	0.0136					5.192	5.682	5.682
1.0	2.914	1.236	0.0758	0.0152	0.0054	0.0019	0.0015		4.250	4.638	4.638
1.1	2.309	1.036	0.0814	0.0142	0.0066	0.0025	0.0011		3.451	3.749	3.749
1.2	1.840	0.8702	0.0850	0.0139	0.0069	0.0031	0.0015		2.821	3.031	3.031
1.5	0.9806	0.5249	0.0863	0.0152	0.0063	0.0031	0.0019		1.618	1.735	1.735
2.0	0.4184	0.2702	0.0828	0.0256	0.0111	0.0030	0.0019		0.8130	0.9045	0.9045

E. $1s$–$2s$–$2p$ eigenstates coupling approximation

	Singlet										
k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L7$	$\sum_S$		
0.9	0.4503	0.0133	0.0457	0.0058	0.0000				0.5151		
1.0	0.2649	0.0100	0.0660	0.0076	0.0019	0.0007	0.0007		0.3518		
1.1	0.1736	0.0131	0.0582	0.0104	0.0028	0.0009	0.0004		0.2594		
1.2	0.1283	0.0102	0.0351	0.0104	0.0030	0.0012	0.0005		0.1887		
1.5	0.0854	0.0021	0.0049	0.0042	0.0025	0.0012	0.0007	0.0010	0.1020		
2.0	0.0591	0.0093	0.0016	0.0009	0.0007	0.0006	0.0004	0.0002	0.0728		
3.0	0.0266	0.0138	0.0052	0.0019	0.0008	0.0003	0.0002	0.0005	0.0493		
4.0	0.0133	0.0106	0.0061	0.0034	0.0020	0.0012			0.0366		
	Triplet										
k_1	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_T$	$\sum_S+\sum_T$	Q_T
0.9	3.687	1.726	0.0875	0.0166	0.0055				5.523	6.037	6.037
1.0	2.902	1.372	0.0938	0.0174	0.0062	0.0021	0.0016		4.401	4.753	4.753
1.1	2.299	1.090	0.0949	0.0166	0.0075	0.0028	0.0012		3.512	3.772	3.772
1.2	1.828	0.8884	0.0956	0.0162	0.0076	0.0033	0.0015		2.841	3.029	3.029
1.5	0.9697	0.5228	0.0911	0.0166	0.0069	0.0033	0.0021	0.0020	1.615	1.716	1.716
2.0	0.4006	0.2433	0.0657	0.0154	0.0044	0.0020	0.0011	0.0008	0.7333	0.8061	0.8061
3.0	0.1127	0.0854	0.0388	0.0156	0.0062	0.0022	0.0009	0.0007	0.2625	0.3118	0.3120
4.0	0.0475	0.0444	0.0276	0.0160	0.0093	0.0053			0.1501	0.1867	0.1888

TABLE IV.—*2s–2s Elastic Cross Sections*

A. Born approximation

k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum$	Q_T
0.24	389.97	15.910	0.27680	0.00319	0.00003	0.00000	0.00000	0.00000	406.16	406.17
0.50	105.52	32.129	5.0152	0.55306	0.04846	0.00347	0.00022	0.00001	143.269	143.276
0.68	47.381	24.746	7.4981	1.7023	0.31602	0.04869	0.00658	0.00078	81.700	81.703
0.83	26.562	17.895	7.5601	2.5080	0.69820	0.16365	0.03393	0.00621	55.427	55.440
1.23	8.2461	7.6846	4.9702	2.7483	1.3500	0.57745	0.22452	0.07803	25.879	25.990
1.80	2.4331	2.7773	2.2763	1.6805	1.1541	0.71994	0.41776	0.22120	11.680	12.105
2.87	0.5245	0.71381	0.68903	0.61123	0.51876	0.41048	0.31076	0.21860	4.0011	4.8280
3.91	0.18794	0.27611	0.28658	0.27339	0.25098	0.21580	0.17946	0.14029	1.8106	2.7417

B. Exchange neglected 1*s*–2*s*–2*p* eigenstates coupling approximation

k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum$	Q_T
0.24	32.02	8.489	42.26						82.77	82.78
0.50	2.041	7.710	23.35	9.004	9.368	7.702	5.870		65.045	65.052
0.68	2.255	8.247	15.16	5.547	4.373	3.755	2.975		42.312	42.316
0.83	1.861	8.206	10.67	4.637	2.792	2.223	1.818		32.207	32.226
1.23	1.716	4.987	5.030	3.321	1.797	1.058	0.7179		18.537	18.726
1.80	1.020	2.134	2.150	1.750	1.275	0.8320	0.5457		9.707	10.352

C. 1*s*–2*s* eigenstates coupling approximation

Singlet

k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_S$
0.24	0.3303	8.196	0.2628	0.0028					8.792
0.50	1.532	10.38	0.0275	0.0048	0.0008	0.0002	0.0249		11.97
0.68	1.115	5.536	1.502	0.1150	0.0087	0.0010	0.0017		8.279
0.83	0.8980	3.512	1.997	0.4303	0.0747	0.0129	0.0032		6.928
1.23	0.5702	1.413	1.236	0.7010	0.3129	0.1228	0.0450		4.401
1.80	0.2825	0.5370	0.5285	0.4193	0.2931	0.1863	0.1110		2.358

Triplet

k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_T$	$\sum_S+\sum_T$	Q_T
0.24	45.94	118.8	7.713	0.0540					172.51	181.30	181.31
0.50	0.2102	34.44	21.05	2.776	0.2521	0.0316	0.0994		58.86	70.83	70.84
0.68	1.366	18.13	12.74	4.059	0.8282	0.1463	0.0442		37.31	45.60	45.60
0.83	2.112	11.65	8.725	3.887	1.230	0.3225	0.0850		28.01	34.94	34.96
1.23	1.811	4.691	4.008	2.585	1.399	0.6637	0.2862		15.44	19.84	20.03
1.80	0.8989	1.735	1.652	1.316	0.9510	0.6359	0.4008		7.590	9.947	10.592

TABLE IV.—Continued

D. $1s$–$2s$–$2p$ eigenstates coupling approximation

	Singlet										
k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_S$		
0.24	7.800	14.79	22.42						45.01		
0.50	0.2858	0.6960	3.139	4.447	2.925	2.063	1.491		15.057		
0.68	0.0661	1.044	2.455	1.884	1.480	1.098	0.7928		8.820		
0.83	0.1675	1.088	2.105	0.9905	0.8282	0.6606	0.5071		6.347		
1.23	0.3739	1.075	1.196	0.7196	0.3940	0.2496	0.1827	0.1409	4.332		
1.80	0.2416	0.4974	0.5133	0.4133	0.2923	0.1928	0.1247	0.0799	2.3553		
2.87	0.0847	0.1489	0.1574	0.1466	0.1280	0.1057	0.0852	0.0658	0.9223		
3.91	0.0365	0.0635	0.0702	0.0698	0.0660	0.0599			0.3659		
	Triplet										
k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_T$	$\sum_S+\sum_T$	Q_T
0.24	16.52[a]	1.236	89.65[a]						107.41	152.42	152.43
0.50	6.172	17.19	30.69	12.90	5.559	5.373	4.257		82.14	97.206	97.213
0.68	2.346	12.88	14.52	7.937	3.046	2.459	2.090		45.28	54.099	54.103
0.83	1.709	9.166	8.976	5.367	2.365	1.514	1.246		30.343	36.686	36.705
1.23	1.391	4.199	3.880	2.681	1.585	0.9110	0.5658	0.4094	15.622	19.955	20.066
1.80	0.7898	1.656	1.603	1.291	0.9466	0.6520	0.4305	0.2786	7.648	10.002	10.427
2.87	0.2622	0.4652	0.4857	0.4479	0.3901	0.3234	0.2623	0.2045	2.8413	3.7636	4.5905
3.91	0.1113	0.1945	0.2146	0.2122	0.2000	0.1812			1.1138	1.4797	2.7306

TABLE V.—*2p–2p Elastic Cross Sections*

A. L–l_1–l_2 odd, Born approximation

k_2	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_o$
0.24	26.562	6.1961	2.1260	0.92680	0.44497	0.22694	0.08105	36.56
0.50	14.210	5.2190	2.1768	1.0526	0.55433	0.31476	0.18383	23.71
0.68	8.8346	4.0761	1.9476	1.0100	0.54983	0.31728	0.18718	16.923
0.83	6.0065	3.2062	1.6993	0.94122	0.53204	0.31331	0.18680	12.8854
1.23	2.5053	1.7048	1.1013	0.70920	0.44518	0.28212	0.17577	6.9237
1.80	0.91169	0.74615	0.56949	0.42475	0.30158	0.21122	0.14245	3.30733
2.87	0.23647	0.22647	0.20035	0.17178	0.13878	0.10926	0.08177	1.16488

B. L–l_1–l_2 odd, exchange neglected $2p$ eigenstates coupling approximations

k_2	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_o$
0.24	61.12	8.444	2.5808	1.1408	0.6200	0.3720	0.2516	74.52
0.50	15.292	6.884	2.5436	1.1580	0.6216	0.3640	0.2352	27.10
0.68	8.008	4.940	2.2548	1.1160	0.6160	0.3640	0.2352	17.54
0.83	5.108	3.6364	1.9292	1.0380	0.5968	0.3592	0.2340	12.90
1.23	2.0484	1.7392	1.1812	0.7652	0.4972	0.3236	0.2208	6.776
1.80	0.7640	0.7220	0.5816	0.4436	0.3312	0.2400	0.1784	3.261
2.87	0.2132	0.2180	0.2004	0.1776	0.1520	0.1272	0.1064	1.195

C. L–l_1–l_2 odd, $2p$ eigenstates coupling approximation

Singlet

k_2	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_{os}$
0.24	2.963	4.161	0.6725	0.2861	0.1552	0.0930	0.0629	8.394
0.50	3.735	3.182	0.7915	0.3066	0.1576	0.0913	0.0589	8.323
0.68	2.165	1.728	0.6851	0.3025	0.1587	0.0920	0.0590	5.190
0.83	1.371	1.107	0.5562	0.2799	0.1547	0.0913	0.0589	3.619
1.23	0.5280	0.4603	0.3115	0.1991	0.1277	0.0823	0.0558	1.7647
1.80	0.1928	0.1831	0.1475	0.1124	0.0837	0.0605	0.0449	0.8249
2.87	0.0534	0.0547	0.0503	0.0445	0.0381	0.0319	0.0267	0.2996

Triplet

k_2	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_{ot}$
0.24	49.22	3.265	1.850	0.8528	0.4650	0.2788	0.1886	56.12
0.50	7.791	1.851	1.504	0.8193	0.4598	0.2722	0.1764	12.87
0.68	4.720	2.219	1.361	0.7700	0.4481	0.2700	0.1756	9.964
0.83	3.373	2.099	1.237	0.7198	0.4318	0.2652	0.1744	8.300
1.23	1.484	1.225	0.8376	0.5509	0.3630	0.2384	0.1639	4.863
1.80	0.5674	0.5337	0.4295	0.3283	0.2456	0.1784	0.1330	2.4159
2.87	0.1596	0.1632	0.1500	0.1328	0.1136	0.0951	0.0797	0.8940

TABLE V—Continued

D. L–l_1–l_2 even, Born approximation

k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_E$	Q_T
0.24	12.488	230.42	4.0427	1.9014	0.30797	0.57480	0.29675	0.06128	250.09	290.69
0.50	0.1758	74.475	7.0963	0.92172	0.48094	0.45745	0.30098	0.05695	83.965	109.714
0.68	0.07386	36.478	6.5573	1.0156	0.29902	0.27895	0.23113	0.04364	44.978	63.742
0.83	0.22032	21.559	5.4143	1.1653	0.26124	0.15835	0.15225	0.02533	28.596	43.572
1.23	0.25024	7.3349	2.9703	1.1453	0.37750	0.11091	0.04377	0.03572	12.269	20.571
1.80	0.13477	2.3518	1.3155	0.74676	0.39290	0.18754	0.07731	0.02824	5.2348	9.6066
2.87	0.04395	0.55809	0.40443	0.30267	0.21866	0.15188	0.09753	0.05832	1.8355	3.8272
3.91	0.01868	0.20876	0.17048	0.14307	0.11656	0.09255	0.06896	0.04864	0.8677	2.1455

E. L–l_1–l_2 even, exchange neglected 1s–2s–2p eigenstates coupling approximation

k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_E$	Q_T
0.24	31.96	91.21	154.6	92.86	55.86	36.79	24.37		487.7	566.32
0.50	9.371	12.65	30.88	20.53	12.99	8.451	5.695		100.6	129.8
0.68	4.156	5.953	13.19	9.528	6.563	4.508	3.130		47.13	66.56
0.83	2.542	4.560	7.424	5.249	3.811	2.793	2.032		28.41	43.07
1.23	1.208	2.879	2.906	1.734	1.106	0.8438	0.6912		11.37	19.56
1.80	0.5612	1.385	1.238	0.8299	0.5109	0.3005	0.2022		5.028	9.382

F. L–l_1–l_2 even, 1s–2p eigenstates coupling approximation

Singlet

k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_{ES}$
0.24	1.964	5.238	17.34	1.260	0.4896	0.2460			26.54
0.50	0.5131	3.159	1.544	1.439	0.4063	0.1906	0.1133		7.365
0.68	0.2346	2.091	1.039	0.4944	0.2539	0.1520	0.0984		4.363
0.83	0.1227	1.252	0.9930	0.1532	0.1191	0.0990	0.0766		2.816
1.23	0.1071	0.6796	0.6167	0.2249	0.0635	0.0249	0.0217		1.7384
1.80	0.0571	0.3366	0.2926	0.1818	0.0954	0.0426	0.0172		1.0233

Triplet

k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_{ET}$	Q_T
0.24	15.62	26.72	32.44	6.674	1.252	0.7359			83.44	178.89
0.50	4.844	3.384	19.97	12.52	0.8285	0.3266	0.2754		42.15	72.80
0.68	2.247	3.886	10.33	5.922	1.035	0.2270	0.1764		23.82	45.22
0.83	1.293	3.650	6.287	3.476	0.9248	0.2097	0.1162		15.957	32.45
1.23	0.4842	2.287	2.424	1.418	0.6072	0.2079	0.0717		7.500	17.28
1.80	0.1909	1.063	0.9673	0.6659	0.4001	0.2024	0.0962		3.586	8.943

TABLE V.—Continued

G. L–l_1–l_2 even, 1s–2s–2p eigenstates coupling approximation

Singlet

k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_{ES}$
0.24	7.852	13.45	38.56	21.41	15.44	8.610	6.432		111.75
0.50	2.470	5.026	7.433	4.900	3.201	2.101	1.417		26.548
0.68	1.344	2.283	3.025	2.683	1.756	1.166	0.7903		13.047
0.83	0.7424	1.316	1.579	1.518	1.079	0.7580	0.5313		7.524
1.23	0.2813	0.6752	0.6254	0.3663	0.2726	0.2289	0.1916		2.641
1.80	0.1357	0.3339	0.2862	0.1816	0.1065	0.0659	0.0471	0.0383	1.1952
2.87	0.0513	0.1053	0.0930	0.0739	0.0553	0.0391	0.0268	0.0181	0.4628
3.91	0.0267	0.0451	0.0422	0.0373	0.0318	0.0261	0.0212	0.0166	0.2470

Triplet

k_2	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$\sum_{ET}$	Q_T
0.24	27.90	75.79	63.10	87.21	41.31	26.12	17.13		338.56	518.92
0.50	4.337	4.540	20.66	21.62	10.38	6.373	4.255		72.17	122.01
0.68	2.823	4.018	10.75	8.986	5.149	3.337	2.302		37.37	67.46
0.83	2.032	3.674	6.400	4.796	2.950	2.013	1.468		23.333	44.54
1.23	0.9741	2.290	2.465	1.599	0.9568	0.6421	0.4924		9.419	20.11
1.80	0.4291	1.055	0.9303	0.6614	0.4150	0.2540	0.1639	0.1180	4.027	9.531
2.87	0.1564	0.3217	0.2907	0.2330	0.1780	0.1291	0.0907	0.0623	1.4619	3.945
3.91	0.0805	0.1366	0.1285	0.1143	0.0982	0.0810	0.0655	0.0524	0.7570	

REFERENCES

1. W. LICHTEN and S. SCHULTZ, *Phys. Rev.* **116,** 1132 (1959).
2. R. F. STEBBINGS, et al., *Phys. Rev.* **119,** 1939 (1960).
3. W. L. FITE, et al., *Phys. Rev.* **112,** 1151 (1958).
4. R. MARRIOTT, *Proc. Phys. Soc.* (*London*) **72,** 121 (1958).
5. K. SMITH, *Phys. Rev.* **120,** 845 (1960).
6. I. C. PERCIVAL and M. J. SEATON, *Proc. Camb. Phil. Soc.* **53,** 654 (1957).
7. P. G. BURKE and K. SMITH, *Rev. Mod. Phys.* **34,** 458 (1962).
8. H. S. W. MASSEY, *Rev. Mod. Phys.* **28,** 199 (1956).
9. S. KHASHBA and H. S. W. MASSEY, *Proc. Phys. Soc.* (*London*) **71,** 574 (1958).
10. W. KOHN, *Phys. Rev.* **74,** 1763 (1948).
11. M. E. Rose, "Elementary Theory of Angular Momentum," Eq. (3.7), J. Wiley, 1957.
12. Reference 11, Eq. (3.16b).
13. Reference 11, Eq. (4.28).
14. Reference 11, Eq. (4.34).
15. Reference 11, Eq. (3.14).
16. E. U. CONDON and G. H. SHORTLEY, "The Theory of Atomic Spectra, Sec. 14³, Cambridge, 1953.
17. B. H. BRANSDEN and J. S. C. MCKEE, *Proc. Phys. Soc.* **A69,** 422 (1956).
18. J. M. BLATT and V. F. WEISSKOPF, "Theoretical Nuclear Physics," Ch. X, J. Wiley, 1960.
19. Equations (2.29) with their present form and without further simplifications are solvable by the computer.
20. Reference 18. Ch. VIII, Eq. (2.7).
21. Reference 8, Table I.
22. Reference 18, Ch. X, Sec. 4.
23. M. J. SEATON, *Proc. Phys. Soc.* **77,** 174 (1961).
24. L. I. SCHIFF,, "Quantum Mechanics," 1st ed., Ch. 4, 1949.
25. Reference 23, Eq. (3.10).

26. K. OMIDVAR, Research Report No. Cx–37, p. 22, Institute of Mathematical Sciences, New York University, 1959.
27. Reference 4. This description differs slightly from the description of reference 26 and the present paper. In reference 4, v: in Eq. (3.5) is set to zero. This makes $B_{ij}{}^{\mu}=0$. Equations (3.6) then reduces to a set of homogeneous equations whose determinant must be zero. Since the amplitude of any of the four components of $\mathbf{u}$ can be left arbitrary, one of the $C_{kl}{}^{\nu}$ is set equal to 1 and the rest of the constants are found subsequently.
28. N. F. MOTT and H. S. W. MASSEY, "The Theory of Atomic Collisions," 2nd ed., Ch. II, 1949.
29. W. E. MILNE, "Numerical Calculus," Sec. 40, Princeton University Press, 1949.
30. Reference 29, Sec. 33.
31. M. J. SEATON, *Peoc. Phys. Soc.* **A68,** 457 (1955).
32. P. G. BURKE, H. M. SCHEY, and K. SMITH, *Phys. Rev.* **129,** 1258 (1963). Similar calculation has been performed by R. Damburg and R. Peterkop which will appear in the USSR Journal of Experimental and Theoretical Physics.

ON THE BREAKDOWN VOLTAGES OF SOME ELECTRONEGATIVE GASES AT LOW PRESSURES*

STEFAN SCHREIER†

Goddard Space Flight Center

The breakdown voltages of five Freon gases and SF_6 were measured at low pressures between parallel plates to determine their insulating properties under these conditions, and especially to find their minimum breakdown voltages. It was found that the advantages of these gases over air are much less at low pressures than at atmospheric pressure or higher and that the use of these gases as insulators at low pressures is therefore limited. Preliminary investigations revealed, however, that the vapors of compounds that are liquid at normal temperature and pressure, such as FC-75, maintain their dielectric strength much better than do electronegative gases as the pressure is reduced, and thus seem to be promising for use as insulators at low pressures.

INTRODUCTION

Certain artificial earth satellites, with direct-current power supplies in the range of 1600 to 2000 volts, have difficulties caused by corona-type discharges at altitudes of about 100,000 feet. These difficulties occur because the equivalent pd (p is the pressure, and d the distance between electrodes) in the satellites passes through the range of minimum breakdown voltage, which in air is about 350 volts. No difficulties occur at either higher or lower altitudes, as might be expected from Paschen's Law, which states that the breakdown voltage between two electrodes is a function of the pressure times the distance between the electrodes and decreases with this product down to a certain minimum, after which it rises again as pd decreases further.

The exact shape and location of the curve of pd versus breakdown voltage will depend on the shape of the electrodes, the material of which they are made, and the nature of the dielectric between them.

Various suggestions for preventing corona-type discharges in the satellites have been made. The solution presently in use consists of imbedding all electrodes (charged conductors) in a solid dielectric so that gaseous discharges are impossible. This is disadvantageous in that conductors, once imbedded, are difficult to remove. Other possibilities include using electronegative gases as insulators (the subject of the present investigation), pressurizing high voltage components, immersing these components in oil, evacuating the entire system, keeping the power off until the critical values of pd have been passed (delayed switching), and artificially suppressing the free electrons in the gas. Some of these suggestions may well bear further investigation. Other possible avenues of investigation are the seeking of improved design criteria to delay corona and the further study of the effects of electrode materials and configurations, not only to prevent, if possible, the onset of corona but also to provide the designer with improved methods of predicting when corona may occur.

To solve the satellite problem, it was proposed to enclose the high voltage components in an electronegative gas that would be allowed to leak out as the satellite rose, so that the pressure inside the satellite would remain about the same as the

* Work performed at Goddard Space Flight Center as part of the 1962 Summer Workshop Program.

† Now at University of Maryland as PhD candidate in Aeronautical Engineering.

pressure outside. This would eliminate the need for pressurization and the danger of unintentional leaks, and would prevent corona-type discharges as long as the minimum breakdown voltage of the gas was higher than the maximum voltage expected in the satellite. For this reason, a study was undertaken to find the minimum breakdown voltages of certain electronegative gases.

PREVIOUS WORK ON ELECTRONEGATIVE GASES

It is well known that electronegative gases obtain their dielectric strength from the fact that their molecules have a tendency to absorb loose electrons and thus inhibit the electron avalanches necessary for breakdown. This already has been discussed briefly by Meek and Craggs in reference 1. A somewhat longer summary is given by Devins and Sharbaugh in reference 2. The effect of pressure on the positive point-to-plane discharge for some electronegative gases was investigated in 1939 by Pollack and Cooper.[3] After the war, further investigations were undertaken by Camilli and Chapman[4] and Wilson et al.[5] Further work was done on fluorine-containing gases by Camilli and Plump in 1952.[6] Additional theoretical considerations were published by Geballe and Reeves in 1953.[7] The electric breakdown on perfluorocarbon vapors and their mixtures with nitrogen was investigated by Berberich et al. in 1955;[8] the dielectric behavior of some fluorogases and their mixtures was also discussed in the same year by Camilli et al.[8] Further study of factors controlling electric strength of gaseous insulation was done by Narbut et al. in 1959.[10] The advantages of gas-insulated power transformers were presented by Camilli in 1959.[11] In the same year Blodgett evaluated some of the dielectric properties of octafluorocyclobutane.[12] A further discussion of fluorocarbons as electrical insulators was given by Reuther in 1951.[13]

The above is by no means a complete listing but is given merely as an illustration of some of the work that has been done in this field. Further references may be found in AIEE Special Publication S-97.[14] The difficulty with most of the previous work is that it was concerned chiefly with the insulating properties of gases at atmosphere or higher pressures; very little has been said about breakdown at lower pressures and especially about minimum breakdown voltages.

GASES CONSIDERED IN THE PRESENT STUDY

The gases considered in the present study were Freons 14, 114, 115, 116, C318, and SF_6. Sulfurhexafluoride (SF_6) was chosen because it is perhaps the best known and most used of the electronegative gases. The newer Freons were evaluated because little is as yet known about their properties and it was hoped that they would prove superior to their older "relatives."

Freon 14 is tetrafluoromethane, CF_4. It has a molecular weight of 88.01, boils at −128°C, and freezes at −184°C at 1 atmosphere pressure. Its dielectric strength at 1 atmosphere and 23°C relative to nitrogen is 1, and its dielectric constant is 1.0006 at 24.5°C and 1 atmosphere.

Freon 114 is dichlorotetrafluoroethane $CC1F_2$—$CC1F_2$. It has a molecular weight of 170.93, boils at 3.55°C, and freezes at −94°C at 1 atmosphere pressure. Its dielectric strength relative to nitrogen is 2.8 at 1 atmosphere and 23°C, and its dielectric constant is 1.0021 at 26.8°C and 1 atmosphere.

Freon 115 is chloropentafluoroethane, $CC1F_2$—CF_4. It has a molecular weight of 154.48, boils at −38°C, and freezes at −106°C at 1 atmosphere pressure. Its dielectric strength relative to nitrogen is 2.8 at 1 atmosphere and 23°C, and its dielectric constant is 1.0018 at 27.4°C and 1 atmosphere.

Freon 116 is hexafluoroethane, C_2F_6. It has a molecular weight of 138.0, boils at −78.2°C and freezes at −100.6°C at 1 atmosphere pressure. Its dielectric strength relative to nitrogen is 1.96 at 1 atmosphere, and its dielectric constant is 1.00197 at 23°C and 711 mm Hg.

Freon C318 is octafluorocyclobutane, C_4F^8 (cyclic). It has a molecular weight of 200, boils at −6.0°C, and freezes at −41.4°C at 1 atmosphere pressure. Its dielectric strength relative to nitrogen is 2.63 at 1 atmosphere, and its dielectric constant is 1.0034 at 10°C and 760 mm Hg.

The properties of *sulfurhexafluoride*, SF_6, are well known. Its molecular weight is 146, it boils at −63.8°C at 1 atmosphere pressure, and its dielectric strength relative to nitrogen is 2.5 at 1 atmosphere.

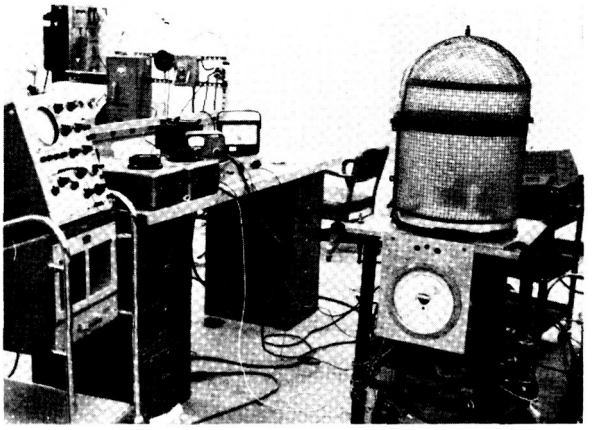


FIGURE 1.—Experimental equipment for determining breakdown voltage of electronegative gases. Total setup showing scope, variacs, vacuum gage, voltmeter, bell jar, and pressure gage.

THE EQUIPMENT USED

The equipment used was a variant of that used by Dakin and Lim[15] of Westinghouse and that of E. L. Brancato of the Naval Research Laboratory (figures 1 and 2). Two parallel circular brass plates, 4 inches in diameter and $\frac{1}{4}$ inch thick with rounded edges of $\frac{1}{8}$ inch radius, mounted on a Teflon base, were used as electrodes. These were connected to a 3000 volt transformer connected to an ordinary 60 cycle, 110 volt power supply. The 60 cycle alternating current, while yielding the same type of corona as direct current, has the advantage over direct current that, because of its relatively slow oscillations, it triggers corona at each peak of the cycle on a regular basis at the proper voltage—and thus makes it much easier to study the phenomenon (figure 3). Filters were used both to suppress high frequency noise from the power supply before it reached the electrodes and to filter out the 60 cycle sine wave before feeding the signal to the oscilloscope. The pressence of corona was determined by observing the signal from the electrodes on a Tektronix type 545A oscilloscope, using a type L plug-in unit preamplifier with a fast rise time. The electrodes themselves were placed in a bell jar that was evacuated by means of a mechanical pump capable of reducing the pressure in the bell jar to 45 microns Hg. Voltages across the electrodes were read from a carefully calibrated Simpson vacuum-tube voltmeter. The pressure inside the bell jar was read from a Wallace and Tierney

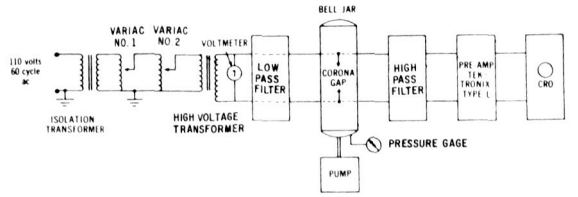
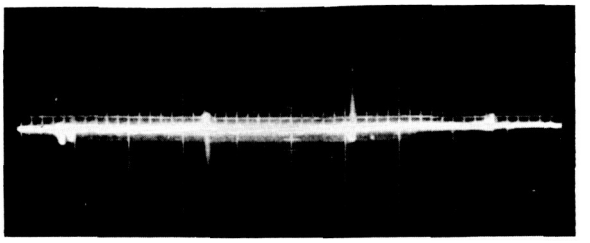
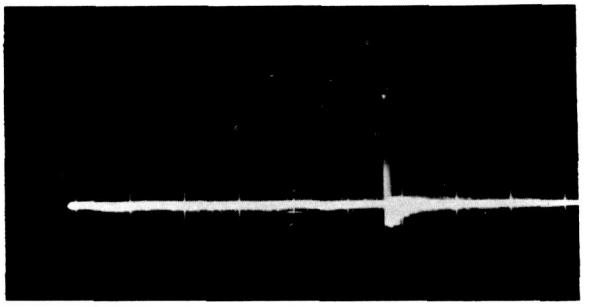


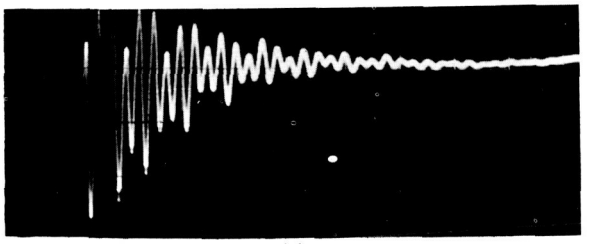
FIGURE 2.—Experimental equipment.



(a)



(b)



(c)

FIGURE 3.—Phenomenon of corona discharge. (a) Corona discharges being triggered by 60 cycle sine wave. (b) Single corona pulse, showing initial spike (double exposure) followed by reverberation. (c) Detail of reverberation following corona pulse (about 1 millisecond).

absolute-pressure gage calibrated to read NACA 1956 standard atmosphere to 200,000 feet; from 1 mm Hg down, the pressure was also read from a Consolidated Vacuum Corporation thermocouple vacuum gage to give a double check in the range of equivalent altitude from about 150,000 to 200,000 feet. The filters were immersed in transformer oil to prevent the possibility of local discharges occurring in the circuit.

THE EXPERIMENTAL PROCEDURE

The equipment was first tested in air (figure 4) to make certain that it was operating properly and to provide a basis for comparison with subsequent readings, as it is well known that the exact shape of the Paschen curve tends to depend on the particular configuration and equipment being used. To test other gases, the bell jar was first evacuated to a pressure of 60 milcrons Hg; and enough gas was introduced to raise the pressure to 60 mm Hg or more—thus guaranteeing a purity of 99.9 percent in the gas being tested. After the pressure was recorded, the voltage was raised across the electrodes until breakdown was observed on the oscilloscope. The voltage was then backed off until the discharge ceased and was raised again in small increments until corona was again observed. This procedure was followed to prevent errors due to time-lag effects. For each reading the entire procedure was followed twice to make sure the reading was correct. Once the reading was obtained for a given pressure, enough gas was pumped out to lower the pressure by about 10,000 feet equivalent altitude (around 5000 feet near the minimum voltage), and the next breakdown voltage was obtained. Three runs were made for each gas, with a gapwidth between the electrodes of 1 mm, $\frac{2}{3}$ mm, and $\frac{1}{3}$ mm.

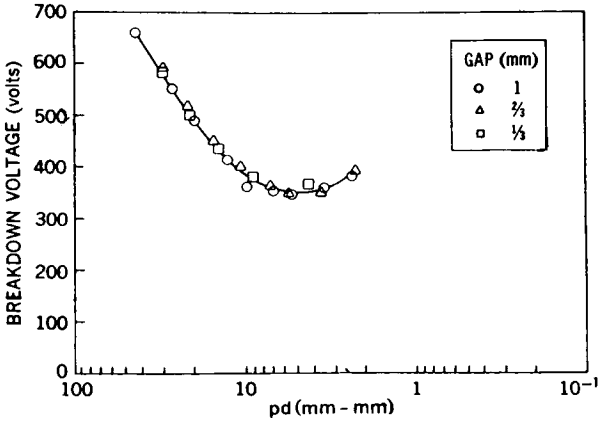


FIGURE 4.—Paschen curve for air test.

CONCLUSIONS

The results show that the electronegative gases tested lose most of their advantage over air as insulators at lower pressures. Freon 114, which has one of the highest dielectric strengths of any of the Freons at atmospheric pressure, had the lowest minimum breakdown voltage of any of the gases tested (about 435 volts). The gas SF_6 had the highest minimum breakdown voltage (520 volts), as compared with a minimum for air of 353 volts, in the present configuration. Note that the relation between breakdown voltage and molecular weight mentioned by Camilli[11] ceases to hold for the minimum breakdown voltage.

The results of these tests, shown in figure 5, indicate that gases of the type tested are not too promising as insulators at lower pressures in satellites. On the other hand, preliminary tests indicate that vapors of such liquids as FC-75 seem to keep their dielectric properties much better than gases at lower pressures. Two new Freon compounds currently under development, which are liquid at normal temperature and pressure, also hold promise; they are 1,1,1,3-tetrachlorofluoropropane, and 1,1,1-trichloropentafluoropropane. The former especially is claimed by the manufacturer to have a dielectric strength in vapor form 4.7 times that of FC-75. These vapors would seem to merit further investigation.

ACKNOWLEDGMENTS

The author is especially indebted to Dr. Elias Klein of Goddard Space Flight Center and to Dr. Donald L. Waidelich of the University of Missouri for their constant assistance and advice during the course of this work. He would also like to express his appreciation to Dr. Thomas W. Dakin of Westinghouse and Mr. G. Camilli, formerly of General Electric, for their helpful advice and to Dr. Eugene Coyner of E. I. Du Pont

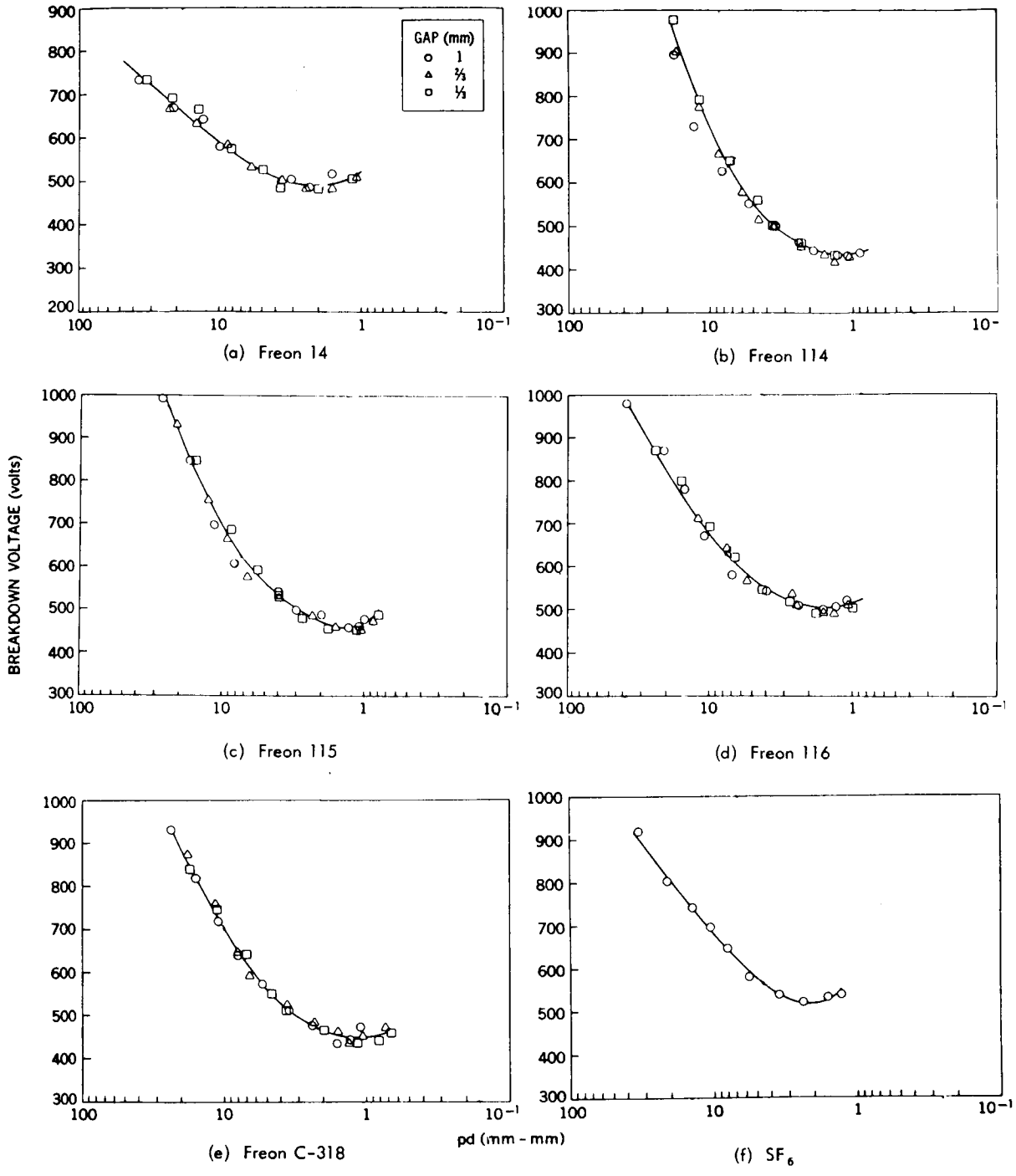


FIGURE 5.—Paschen curves for Freon gases and SFi_6.

de Nemours and Company, who supplied the Freon gases used and who gave generously of his time and counsel.

Last but not least, the author would like to express his appreciation for the assistance given him by the many technicians and machinists of the Goddard Space Flight Center, without whose assistance this work would not have been possible.

REFERENCES

1. Meek, J. M., and Craggs, J. D., "Electrical Breakdown of Gases," Oxford: Clarendon Press, 1953.
2. Devins, J. C., and Sharbaugh, A. H., "The Fundamental Nature of Electrical Breakdown," *Electro. Tech.* **67**(2): 104–122, February 1961.
3. Pollack, H. D., and Cooper, F. S., "The Effect of Pressure on the Positive Point-to-Plane Discharge in N_2, O_2, CO_2, SO_2, SF_6, CCl_2F_2, A, He, and H_2," *Phys. Rev.* **56**(2): 170–175, July 15, 1939.
4. Camilli, G., and Chapman, J. J., "Gaseous Insulation for High-Voltage Apparatus," *Trans. Amer. Inst. Elect. Engrs.* 66: 1463–1470, 1947.
5. Wilson, W. A., Simons, J. H., and Brice, T. J., "The Dielectric Strength of Gaseous Fluorocarbons," *J. Appl. Phys.* **21**(3): 203–205, March 1950.
6. Camilli, G., and Plump, R. E., "Fluorine-Containing Gaseous Dielectrics," *Trans. Amer. Inst. Elect. Engrs.* 72, Pt. 1 (*Comm. and Electronics*) No. 6: 93–102, May 1953.
7. Geballe, R., and Reeves, M. L., "A Condition on Uniform Field Breakdown in Electron-Attaching Gases," *Phys. Rev.* **92**(4): 867–868, November 15, 1953.
8. Berberich, L. J., Works, C. N., and Lindsay, E. W., "Electric Breakdown of Perfluorocarbon Vapors and Their Mixtures With Nitrogen," *Trans. Amer. Inst. Elect. Engrs.* 74, Pt. 1 (*Comm. and Electronics*) No. 21: 660–666, November 1955.
9. Camilli, G., Liao, T. W., and Plump, R. E., "The Dielectric Behavior of Some Fluorogases and Their Mixtures," *Trans. Amer. Inst. Elect. Engrs.* 74, Pt. 1 (*Comm. and Electronics*) No. 21: 637–641, November 1955.
10. Narbut, P., Berg, D., et al., "Factors Controlling Electric Strength of Gaseous Insulation," *Trans. Amer. Inst. Elect. Engrs.* 78, Pt. 3 (*Power Apparatus and Systems*) No. 43: 545–551, August, 1959.
11. Camilli, G., "Gas-Insulated Power Transformers," *Proc. Inst. Elect. Engrs.* **107A**(34): 375–382, August 1960.
12. Blodgett, F. W., "Properties of Octafluorocyclobutane, a Dielectric Gas," *Trans. Amer. Inst. Elect. Engrs.* 78, Pt. 1 (*Comm. and Electronics*) No. 41: 63–66, March 1959.
13. Reuther, H., "Fluorkarbone als Elektrische Isolierstoffe," *Elektrie* **15**(6): 184–186, June 1961.
14. "Bibliography on Gaseous Dielectric Phenomena," *Amer. Inst. Elect. Engrs. Spec. Publ.* S-97, April 1957.
15. Dakin, T. W., and Lim, J., "Corona Measurement and Interpretation," *Trans. Amer. Inst. Elect. Engrs.* 76, Pt. 3 (*Power Apparatus and Systems*) No. 33: 1059–1065, December 1957.

THE STRUCTURE OF ZINC OXIDE NUCLEI

G. W. SEARS

Research Division, General Dynamics/Electronics

AND

R. POWELL AND B. DONN

Goddard Space Flight Center

INTRODUCTION

A common morphology of zinc oxide particles formed by oxidation of zinc vapor consists of four acicular crystals united at a common juncture. Such particles are called fourlings. Fuller[1] has made a crystallographic study of fourlings with the aid of stereo-electron microscopy. He has determined that the crystallographic *c*-axis coincides with the axis of each leg. The four crystals are twinned on planes of type (112).

In the classical analysis[2,3] of the nucleation of a crystalline phase from a parent fluid phase it is assumed that the nuclei are tiny perfect crystals of the forming phase. This analytical assumption has become so identified with the structure of real nuclei that the idea of imperfect nuclei has been seriously discussed[4] in terms of empirical observations only recently. One of the past difficulties has been in ascertaining the structure of the nucleus.

It is the purpose of the present paper to demonstrate that the fourfold twin structure of zinc oxide fourlings are representative of the structure of the original nuclei. In brief the fourlings nucleate as fourfold twins, a poly-crystalline structure as opposed to the assumption of classical nucleation of a single crystal nucleus. Although it is possible to establish growth conditions shortly after nucleation on the basis of growth features, it is not possible to deduce the degree of supersaturation at the time of nucleation. Alternative rationalizations of the nucleation event are presented.

EXPERIMENTAL

The zinc oxide particles were grown by passing a mixture of oxygen and a diluent gas over a container of molten zinc heated to about 900°C with an electrical resistance furnace. The furnace and container were an integral unit supported inside a quartz reaction vessel. The ratio of oxygen to diluent gas was established by monitoring the flow rates independently prior to mixing.

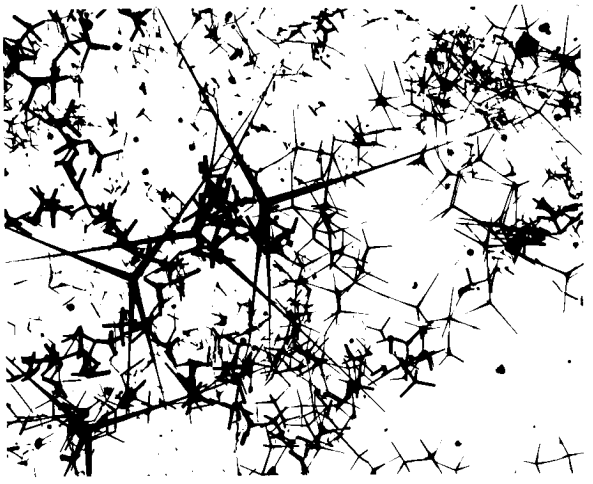


FIGURE 1.—ZnO nuclei.

The smoke particles were deposited from the gas stream upon a thin carbon film. The film with particles was mounted on an electron microscope grid and examined. Fig. 1 shows a typical field at 40,000× magnifications.

Three main features pertinent to our interpretations are evident. (1) Fourlings have been observed in which the central junction ranges from a diameter of as large as 1000 Å to as small as 30 Å. The total diameter of the tiniest fourlings including the legs was as small as 300 Å. (2) In many cases the legs grew at uniform cross-sections as crystalline whiskers. (3) The number density of fourlings in a typical smoke was many orders of magnitude higher than the concentration of foreign dust particles in the input gas.[4] Frequently very thin platelets bounded by *c*-surfaces are observed in the same field as fourlings. There is some evidence that the shape of the particle depended upon the carrier gas, i.e., a larger proportion of fourlings appeared to be obtained with N_2 compared with He or Ar as carrier.

DISCUSSION

Formation of Fourlings

In a study of the nucleation of alumina at high supersaturations Sears and DeVries[5] have described the necessary conditions for deducing the structure of a nucleus from observations on a crystallite grown therefrom. A given nuclei should form from a limited volume of sufficiently supersaturated parent phase. The supersaturation should rapidly decrease so that subsequent growth occurs at a relatively low supersaturation.

In the present experiments the supersaturation required for three-dimensional nucleation is not known. However the final morphology shows that in may cases subsequent growth took place at a relatively low supersaturation. Many fourlings have arms of quite uniform cross-section and morphologically can be described as whiskers. In principle they could have grown if the local supersaturation was insufficient to cause two-dimensional nucleation on columnar surfaces and was supercritical for two-dimensional nucleation on the *c*-surfaces. However, the existence of thin platelets bounded by *c*-surfaces strongly indicates that the local supersaturation after nucleation was insufficient to cause two-dimensional nucleation of the *c*-surfaces. If true, the *c*-whiskers must have grown by defect mechanism,[6] most probably by a screw dislocation.

In summary the model of a fourling is a fourfold twin structure with (112) twinning planes and with a screw dislocation in the *c*-direction of each twin. The existence of the dislocations is compatible with Fuller's observation[1] that a distortion must exist at the juncture of each pair of legs of the fourling.

From the preceding discussion it has been established that, after nucleation had occurred at some undetermined supersaturation, subsequent growth occurred at a relatively low supersaturation. These are the conditions[5] that justify identifying a crystallite structure with that of the parent nucleus. The additional argument can be made that the fourling structure is invarient over a range of central junction diameter from 1000 Å to ~30Å. It might be suspected that the nucleus diameter might not be a great deal smaller, if it is to contain four twins and their twinning boundaries.

The initiation of fourlings must occur by homogeneous nucleation. The ratio of fourling particles in a fixed volume of smoke is many powers of ten greater than the number of foreign particles, or potential nucleation heterogeneities, in a like volume of input gas.

APPROPRIATE NUCLEATION THEORY

Classical nucleation theory[2,3] has been derived from Gibb's theory[7] of phase stability. Gibbs recognized that a small particle of bulk phase might be completely inhomogeneous in composition. However, he developed a self-consistent description of a crystalline particle by treating it as though it was homogeneous up to a sharp interface with its environmental phase. Nucleation theory adopted Gibb's homogeneous particle description. The volume free energy was equated to a value characteristic of the bulk phase and the specific interfacial free energy was equated to that of a large flat interface. In some manner the model has become identified with reality as far as the homogeneity of the nucleus is concerned.

Turnbull[8] recognized certain inadequacies in classical theory and stated, "The free energy of a cluster may be a minimum for a molecular arrangement quite different from that of the bulk phase at equilibrium." However, it remained for Cahn and Hilliard[9] to develop a satisfactory analysis of nucleation energetics based upon their thermo-

dynamic treatment of non-uniform systems.[10] They were able to evaluate the excess free energy associated with the formation of a new phase without an arbitrary assignment of surface and volume free energy terms. The compositional variation across a nucleus is the one yielding the lowest free energy barrier to nucleation while maintaining the average composition of the system invariant. By including gradient free energy terms, the interfacial free energy is automatically included. In summary, any assumptions as to the morphology and homogeneity of the nucleus are avoided.

The Cahn–Hilliard treatment agrees with classical theory at low supersaturations. With increasing supersaturation the radius of a critical nucleus reaches a minimum and increases without bound thereafter. With increasing supersaturation a nucleus becomes less and less homogeneous. Nucleation in the regieme of inhomogeneous nuclei is called non-classical nucleation.

Sears[4] has discussed the conditions under which non-classical nucleation of crystalline solids should be expected. It was concluded that the conditions for non-classical nucleation can be attained. These conditions are normally associated with either spherulite or glass formation. It was proposed that inhomogeneity of a nucleus corresponds to either a polycrystalline or spherulitic nucleus, or to a rigid glass.

One possible description of the nucleation of zinc oxide fourlings is that the event occurs at high supersaturations in the non-classical regieme.

However an analysis by Fullman[11] of the equilibrium form of crystalline bodies indicates a second alternative. He concluded that the lowest free energy form of a given mass of crystalline phase might be as a polycrystalline body. It was required that the crystal have an anisotropic surface energy. Under this condition it is possible that the total interfacial and surface free of a polycrystalline body is less than the minimal free energy of bounding the same material as a single crystal.

Although Fullman could show no demonstrable examples, it is entirely possible that a fourfold twin of zinc oxide bounded by *c*-surfaces and containing twin planes might be more stable than a single crystal in its equilibrium form. A final condition must be the maintenance of relative stability of the four-fold twin as its dimensions approach those of a nucleus.

The Fullman analysis provides a second alternative for fourling nucleation. This possibility can be distinguished in principle from the non-classical model since it requires a relatively low supersaturation.

CONCLUSION

Zinc oxide fourlings are shown to originate in a single homogeneous nucleation event. Alternative descriptions of the nucleation process are outlined. The appropriate description requires knowledge of the critical supersaturation during nucleation. The experimental procedures used for examination of zinc oxide smoke should provide a quite general method of deducing the structure of a variety of nuclei.

ACKNOWLEDGMENT

The work of G. W. Sears was carried out under Contract NAS5-1590 between General Dynamics/Electronics/Rochester, and Goddard Space Flight Center.

REFERENCES

1. M. L. Fuller, *J. Appl. Phys.*, **15,** 164 (1944).
2. M. Volmer and A. Weber, *Z. Physik. Chem.*, **119,** 277 (1925).
3. R. Becker and W. Döring, *Ann. Physik* (5) **24,** 719 (1935).
4. G. W. Sears, *J. Phys. Chem.*, **65,** 1738 (1961).
5. G. W. Sears and R. C. DeVries, *J. Chem. Phys.*, **32,** 93 (1960).
6. G. W. Sears, *Acta Met.*, **1,** 457 (1953).
7. J. W. Gibbs, "Collected Works," Yale University Press, New Haven, Conn., 1948, Vol. I, pp. 105–115, pp. 252–258.
8. D. Turnbull in Solid State Physics, Ed. by F. Seitz and D. Turnbull, Academic Press, Inc., New York, 1956, Vol. 3, p. 261.
9. J. W. Cahn and J. E. Hilliard, *J. Chem. Phys.*, **31,** 688 (1959).
10. J. W. Cahn and J. E. Hilliard, *J. Chem. Phys.*, **28,** 258 (1958).
11. R. L. Fullman, *Acta Met.*, **5,** 638 (1957).

ELECTRON-HYDROGEN PHASE SHIFTS JUST BELOW THE INELASTIC THRESHOLD

AARON TEMKIN

Goddard Space Flight Center

The singlet S-wave phase shifts for the scattering of electrons by atomic hydrogen just below the inelastic threshold (10.2 ev) are calculated by using a recently developed nonadiabatic theory. Previous calculations of Burke and Schey based on the much more restricted close-coupling approximation yielded a resonance characterized by the phase shift going through π radians. In view of a theorem of Wigner and a calculation of Holøien this result is indicative of a bound state of the compound system near the energy 9.4 ev. It is argued on the basis of a previous calculation that this resonance must be distinct from another resonance near 10.1 ev. And on this basis it is concluded that this partial cross-section must exhibit two resonances below threshold.

INTRODUCTION

Burke and Schey[1] have recently calculated a resonance in the elastic scattering of electrons by atomic hydrogen just below the threshold for inelastic scattering (10.203 ev). The basis of their calculation was the close-coupling approximation with $1s$–$2s$–$2p$ states. The nomenclature refers to hydrogenic states, each of which is multiplied by an initially undetermined function. For S-wave† scattering, their wave function can be written:

$$r_1 r_2 \Psi_{cc} = \left[u(r_1) R_{1s}(r_2) + v(r_1) R_{2s}(r_2) + (1 \rightleftarrows 2)\right] P_0(\cos\theta_{12}) + \sqrt{3}\left[w(r_1) R_{2p}(r_2) + (1 \rightleftarrows 2)\right] P_1(\cos\theta_{12}) , \quad (1)$$

where $R_{nl}(r)$ is r times the nl radial wave function of hydrogen. It is clear from equation 1 that this function is approximate in two ways. First, it contains only 2 of an infinite number of *relative* angular momenta $P_l(\cos\theta_{12})$. Second, the "coefficients" of the included P_l have a comparatively restricted form (which, however, is manifestly symmetric with respect to the interchange of 1 and 2 corresponding to the necessary symmetry of singlet scattering, with which we shall here be concerned).

Clearly the most general function containing P_0 and P_1 can be written:

$$r_1 r_2 \Psi = \Phi_0(r_1 r_2) P_0(\cos\theta_{12}) + \sqrt{3}\,\Phi_1(r_1 r_2) P_1(\cos\theta_{12}) , \quad (2)$$

where the two-dimensional functions Φ_0 and Φ_1 are required to have the correct symmetry. Substitution of equation 2 into the variational principle, $\delta\int\Psi^*(H-E)\Psi\, d\tau = 0$, yields the following coupled set of partial

† Letters referring to angular momentum states are capitalized when the total angular momentum state (partial wave) is being considered. Lower case l refers to the relative angular momentum within a given total angular momentum state.

TABLE 1.—*Zeroth Order Phases Shifts Below Thresholds*

$k\,(\mathrm{ryd}^{1/2})$	k^2 (impacting energy in ev)	δ_0 (radians)
0.85	9.830	0.721
0.855	9.944	0.729
0.860	10.061	0.7545
0.861	10.085	0.768
0.862	10.108	0.793
0.863	10.132	0.853
0.864	10.155	1.25
0.8645	10.169	3.14
0.865	10.180	3.628
0.8656	10.194	3.738
0.866	10.203	3.781

differential equations* in the region $r_1 \geq r_2$ with appropriate boundary conditions:[2,3]

$$\left(\Delta_{12} + \frac{2}{r_2} + E\right)\Phi_0 = \frac{2r_2}{\sqrt{3}\, r_1^2}\,\Phi_1 , \tag{3a}$$

$$\left[\Delta_{12} - 2\left(\frac{1}{r_1^2} + \frac{1}{r_2^2}\right) + \frac{2}{r_2} - \frac{4r_2^2}{5r_1^2} + E\right]\Phi_1 = \frac{2r_2}{\sqrt{3}\, r_1^2}\,\Phi_0 . \tag{3b}$$

It is the object of the present nonadiabatic theory[2,3] to identify and directly attack these partial differential equations, and thus avoid the second category of approximation implicit in the close-coupling method. The purpose of this note is to report on results of calculations on the lowest order approximation of the equations and to give some speculations concerning the next higher correction.

The lowest order approximation is the so-called zeroth order problem and is given by the equation:

$$\left(\Delta_{12} + \frac{2}{r_2} + E\right)\Phi_0^{(0)} = 0 . \tag{4}$$

For future reference note that this basic relation can be derived:

$$\sin\left(\delta_{(1)} - \delta_0\right) = \frac{-2}{k\sqrt{3}}\int_0^{\infty} dr_1 \int_0^{r_1} dr_2\, \Phi_0^{(0)} \frac{r_2}{r_1^2}\,\Phi_1 , \tag{5}$$

where δ_0 is the phase shift of equation 4, and $\delta_{(1)}$ is the phase shift of equations 3.

In practice, equation 4 has been solved by expanding in exact separable solutions,[3] the program for which has been extended to double precision arithmetic (16 significant figures). For this discussion the resultant values for $\Phi_0^{(0)}$ and δ_0 can be considered correct to the significance given in table 1. A close-coupling approximation of equation 4 gives results in good accord with δ_0.[4]

---

* Energy is in rydbergs (1 ryd = 13.605 ev) and length is in Bohr radii.

CONJECTURES ON THE SOLUTION OF EQUATIONS 3*

An obvious first step in an iterative attempt to solve equations 3 would be to use the $\Phi_0^{(0)}$ in the right-hand side of equation 3b to get an approximate Φ_1 which we can call $\Phi_1^{(0)}$:

$$\left[\Delta_{12} - 2\left(\frac{1}{r_1^2} + \frac{1}{r_2^2}\right) + \frac{2}{r_2} - \frac{4r_2^2}{5r_1^3} + E\right]\Phi_1^{(0)} = \frac{2r_2}{\sqrt{3}\, r_1^2}\Phi_0^{(0)} \,. \tag{6}$$

This neglects the coupling of Φ_0 to Φ_1 (but not the reverse). Figure 1 gives the close-coupling results, labeled δ_{cc}, as well as the results for δ_0. A possible explanation for the shift in energy at which the resonance in δ_{cc} occurs as compared with δ_0 is that it is due to the coupling of Φ_0 to Φ_1, which is included in some approximate way in the close-coupling approximation. However, the previous calculation[5] was based on the assumption that all resonances arise directly from the resonance in $\Phi_0^{(0)}$. The calculated results give absolutely no indication of such a shift. The reason for taking the close coupling seriously is explained below. If it is assumed that there must be a resonance at or below the energy corresponding to the close-coupling resonance, a more plausible explanation of the lower resonance would seem to be that it is due to a resonance (change in sign) of Φ_1 in equation 3b which is fairly independent of the right-hand side.† If this is so then a resonance very close to $k=0.84$ should also come from equation 6. Hopefully, solutions of this equation will be obtained by using a noniterative numerical technique for solving partial differential equations.[7] As another test of this hypothesis a close-coupling expansion with only $1s$–$2p$ states would also be expected to show a resonance near the lower energy.‡

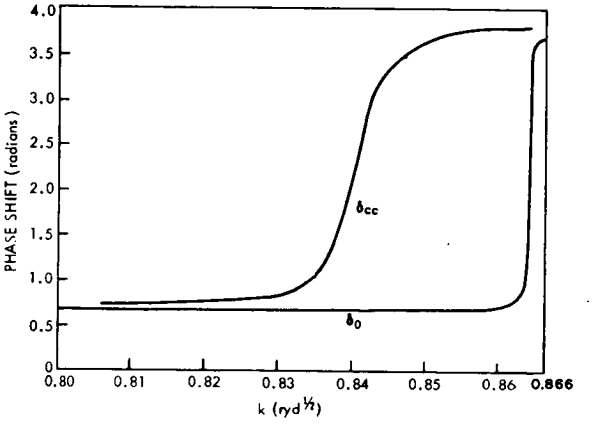


FIGURE 1.—The $1s$–$2s$–$2p$ close-coupling phase shift (δ_{cc}) and δ_0 as a function of k.

Mathematically a resonance in $\Phi_1^{(0)}$ manifests itself in $\delta_1^{(0)}$, the first approximation of $\delta_{(1)}$, whereby the expression

$$\sin\left(\delta_1^{(0)} - \delta_0\right) = \frac{-2}{k\sqrt{3}}\int_0^\infty dr_1 \int_0^{r_1} dr_2\, \Phi_0^{(0)} \frac{r_2}{r_1^2}\Phi_1^{(0)} \,, \tag{7}$$

* The quantitative speculations of this section are quite different from what was calculated in reference 5. The author is particularly indebted to Dr. T. F. O'Malley for pointing out the usual increase of π of phase shifts at resonances and the analytical derivation of this behavior for some simple nonlocal potentials. This discussion led, directly, to a re-examination of δ_0 (reference 5) and the discovery that it does indeed go through π. Dr. O'Malley also brought reference 6 to the author's attention.

† Mr. John Cooper pointed out the importance of the energy shift of the lower resonance.

‡ Schey and Burke are now examining this point.

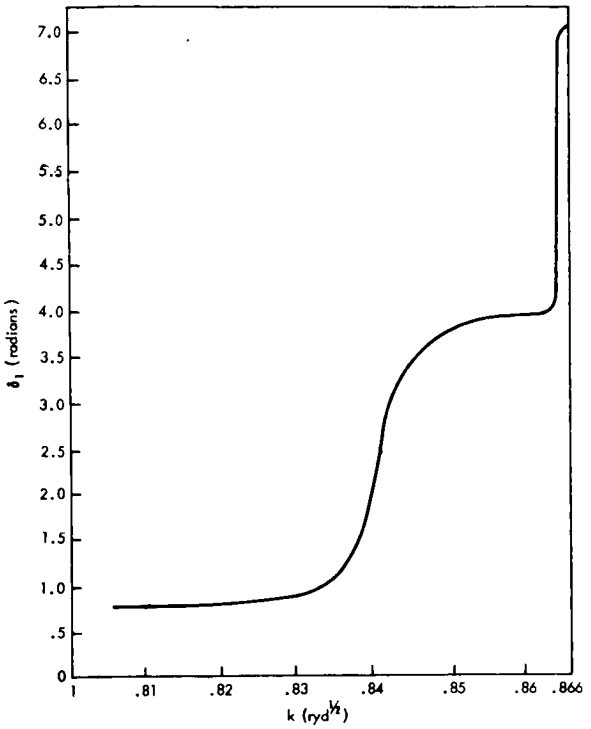


FIGURE 2.—Conjecture of the first order (equations 3) phase shifts according to the arguments in the text.

a change in sign of $\Phi_1^{(0)}$ would mean a change in sign of the integral, since $\Phi_0^{(0)}$ is varying very slowly at the lower energy. This *could* (see below) mean an increase of $\delta_1^{(0)}$ by π which would be the mathematical statement of the resonance.

Along this mathematical line, it might be expected that as the energy increases toward $k=0.8645$, $\Phi_0^{(0)}$ would undergo a change of sign and that there the integral would again change sign, corresponding to an increase of δ_0 into the same quadrant as $\delta_1^{(0)}$. This would be in accord with $\delta_1^{(0)}$ exhibiting only one increase of π corresponding precisely to what Burke and Schey calculated. *However*, here equation 6 must be considered more carefully; it has been presumed to have a resonance due to the structure of the left-hand side at the lower energy. At the higher energy, however, the right-hand side has been shown here to exhibit a resonance and this could cause a second resonance in $\Phi_1^{(0)}$. Mathematically, this would mean that at the same energy at which $\Phi_0^{(0)}$ undergoes a change of sign $\Phi_1^{(0)}$ undergoes a second change in sign. Thus, the integral on the right-hand side of equation 7 undergoes no change of sign because of the simultaneous changes in sign of $\Phi_0^{(0)}$ and $\Phi_1^{(0)}$. This could correspond to a second increase of $\delta_1^{(0)}$ by π. It is problematical whether or not this second phenomenon will correspond, in fact, to a second increase of the phase shift by π (see below). However, it is quite likely that the phase shift will have a sufficiently uneven behavior to appear in the cross-section as a resonance. (It is only a matter of definition whether a resonance is called a sudden increase of the phase shift by π or a sudden up and down behavior in the cross-section.)

In figure 2 the conjectured behavior of $\delta_1^{(0)}$ showing two increases of π is given. The actual values were gotten by superposing the δ_0 resonance (modulo π) on δ_{cc} at the second resonant energy. In figure 2 it has been assumed that resonances correspond to increases of the phase shifts by π (in the case of δ_0 this has been calculated to be the case). This is based on a theorem of Wigner and Eisenbud[6] which says

$$d\delta/dk \geq -R.$$

This statement is derived from the fact that $2(d\delta/dk)$ is the delay time of a particle striking a target. Although there is no upper limit on this quantity, its negative value, which is the time advancement of the scattered wave, is optimal if there is a hard core in the region of interaction. Thus, any reasonable process of the formation of the compound state corresponding to a sticking of the incoming particle to the atom, the phase shift must exhibit a sharp rise. A calculation by Holøien gives an upper limit of $E=9.4$ ev on such a state of the compound H^- ion.[8] A more direct argument that there must be a resonance at or below the close-coupling resonance is afforded by theorems on bounds for phase shifts which have recently been proved.[9]* As it applies in the present case it says that $\delta_1 \geq \delta_0$ up to the inelastic threshold.†

* Burke and Schey have obtained very narrow resonances (private communication) at low energies (≈ 1 eV), whereas careful calculations by Schwartz (private communication) show no resonances. This is not necessarily a contradiction to the theorem of Hahn, O'Malley, and Spruch (reference 9), since the numerical method of Burke and Schey may occasionally show some aberrations. However, other logical possibilities, including a contradiction to the theorem in reference 3, cannot be dismissed at this point.

† Private communication from T. F. O'Malley.

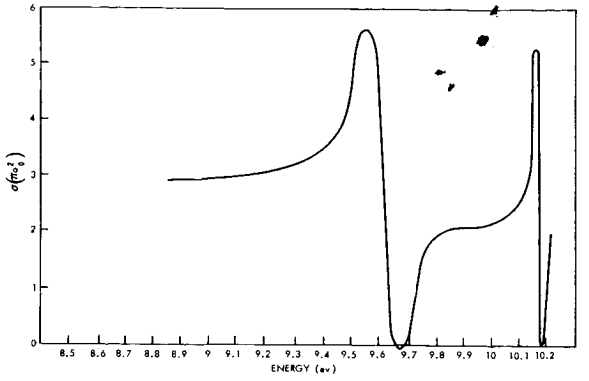


FIGURE 3.—Singlet S–wave cross-section corresponding to the phase shifts in figure 2. The value a_0 is the Bohr radius.

Figure 3 shows the plot of the cross-section corresponding to the phase shift curve in figure 2. This is meant to apply only to the S-wave part of the singlet scattering. In view of the fact that the cross-section contains two resonances, it is felt that the physical reasons for such a behavior which were included in reference 5 still apply. That argument assumes a correspondence of each state of the target with a state of the compound system (with possible shifts in energy). The main reason for thinking that this correspondence does hold in this case is the large energy difference of the lower resonance from the upper one. The previous calculation[5] indicates that otherwise this energy shift would be much smaller.

However, the present heuristic considerations give a more explicit indication of what the physical argument means in terms of the non-adiabatic formalism. The $\Phi_0^{(0)}$ equation (4) is the two-dimensional analogue of the one-particle $2s$ equation. To say that the one-particle $2s$ equation corresponds to a resonance in the compound system means that the $\Phi_0^{(0)}$ inherently contains a spontaneous resonance near the $2s$ energy. The present calculation for δ_0 explicitly yields this resonance. By the same token the left-hand side of equation 3b or equation 6 is a two-particle analogue of the single particle $2p$ equation; therefore, by the same kind of correspondence, this left-hand side is expected also to contain a spontaneous resonance. It should be emphasized that this latter resonance has not as yet been found in the author's calculations. However, the author believes that Burke and Schey have found it,[1] but that they misinterpreted it to be the shifted $\Phi_0^{(0)}$ resonance. The author further believes that if the $1s$–$2s$–$2p$ close-coupling calculation is extended even closer to the $2s$ threshold, it will reveal the second resonance.

It may be inquired whether or not the higher L equations, of which equations 3 are the first two, might also reveal spontaneous resonances. They probably do; however, the one-particle target state energies to which they correspond are higher than p states. Therefore, their principal quantum number is $n \geq 3$ and, thus, their energies should be significantly above the $n=2$ threshold.

REFERENCES

1. BURKE, P. G., and SCHEY, H. M., "Elastic Scattering of Low-Energy Electrons by Atomic Hydrogen," *Phys. Rev.* **126**(1): 147–162, April 1, 1962.
2. TEMKIN, A., "Nonadiabatic Theory of the Scattering of Electrons from Hydrogen," *Phys. Rev. Letters* **4**(11): 566–568, June 1, 1960.
3. TEMKIN, A., "Nonadiabatic Theory of Electron-Hydrogen Scattering," *Phys. Rev.* **126**(1): 130–142, April 1, 1962; also NASA Technical Note D-1343.
4. SMITH, K., MCEACHRAN, R. P., and FRASER, P. A., "Effect of Virtual Excitation of the $2s$ State on the Elastic Scattering of Electrons by Atomic Hydrogen," *Phys. Rev.* **125**(2): 553–558, January 15, 1962.
5. TEMKIN, A., and POHLE, R., "Electron-Hydrogen Phase Shifts Just Below the Inelastic Threshold," *Phys. Rev. Letters* **10**(1): 22–25, January 1, 1963.
6. WIGNER, E. P., "Lower Limit for the Energy Derivative of the Scattering Phase Shift," *Phys. Rev.* **98**(1): 145–147, April 1, 1955.
7. TEMKIN, A., and SULLIVAN, E., "Nonadiabatic Theory of Electron-Hydrogen Scattering. II," *Phys. Rev.* (to be published); also NASA Technical Note D-1702 (to be published).
8. HØLOIEN, E., "The $(2s)^2\ ^1S$ State Solution of the Non-relativistic Schrödinger Equation for Helium and the Negative Hydrogen Ion," *Proc. Phys. Soc. (London)* **71**(3): 357–368, March 1958.
9. HAHN, Y., O'MALLEY, T. F., and SPRUCH, L., "Static Approximations and Bounds on Single Channel Phase Shifts," *Phys. Rev.* **128**(2): 932–943, October 15, 1962.

POSITRON-HYDROGEN SCATTERING

AARON TEMKIN

Goddard Space Flight Center

The nonadiabatic theory of electron-hydrogen scattering is applied to the s-wave scattering of positrons by atomic hydrogen. This zeroth order problem is considerably more difficult than the corresponding one for electron-hydrogen scattering. If we assume that the most extensive of Rotenberg's calculation relating strictly to the zeroth order problem is an adequate approximation, then the present calculations support the conclusion that his final phase shifts are more correct than those of Schwartz. The nonadiabatic theory is also developed for higher partial wave scattering. The portion of the coupled set of equations which gives the major contribution to the phase shifts can readily be isolated. A first approximation of the truncated p-wave equations is obtained; the corresponding phase shifts are found to be positive, but much smaller than those of Bransden.

INTRODUCTION

The obtaining of reliable positron-hydrogen phase shifts has proven to be a very elusive task in spite of the fact that rigorous upper bounds exist for the scattering length[1,2] and in spite of the fact that in the latter reference as many as 50 variational parameters were used in obtaining the scattering length and phase shifts. The uncertainty in Schwartz's work has resulted from a recent calculation of Rotenberg[3] using a novel expansion in terms of Sturmian functions. His phase shifts are considerably larger than those of Schwartz. Further discussion of these differing results will be given later in this paper.

It should be noted, however, the conflicting results notwithstanding, that the phenomenological aspects of this problem are clear (sketched in figure 1). At close distances the positron feels the repulsion of the positive nucleus and at far distances the attraction of the induced dipole moment of the hydrogen atom. These two extremes must be connected in some way. In figure 1 the overall potential (solid curve) is taken to be the sum of the two dashed curves, one representing the repulsive Hartree potential

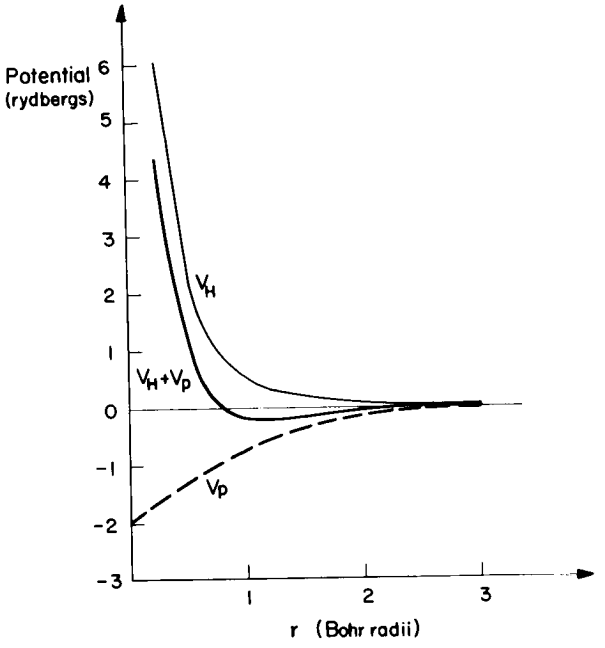


FIGURE 1.—Phenomenological potentials describing the scattering of positions from hydrogen.

and the other representing a phenomenological polarization potential

$$V_H = 2e^{-2r}(1 + r^{-1})$$

$$V_p = \frac{-a}{(r^2 + d^2)^2} ,$$

TABLE 1.—*s-Wave Phase Shifts in Various Calculations*

k (ryd $^{1/2}$)	Phase Shift (radians)				
	d = ∞	d = 1.225	Reference 2	Reference 3	Nonadiabatic†
0*	0.58		−2.10	<−2.7	
0.1	−0.0580	0.21	0.151	0.180	
0.2	−0.1145	0.29	0.188	0.232	0.22
0.3	−0.168	0.31	0.168	0.179	
0.4	−0.2181	0.29	0.120	0.129	
0.5	−0.2636	0.26	0.062	0.068	
0.6	−0.3043	0.23	0.007	0.010	
0.7	−0.340	0.19	−0.054	−0.04	

*The k = 0 entries are scattering lengths.
†Including relative s- and p-waves only.

where $\alpha=4.5d_0^3$ is the well-known polarizability of the hydrogen atom, and d is an adjustable parameter. In Table 1 we have given the phase shifts as a function of k for two values of d, together with the phase shifts of Schwartz[2] and Rotenberg.[3] At $k=0.2$ we have included the result of the present calculation. The column for $d=\infty$ corresponds to the absence of polarization potential. The phase shifts of this static approximation have been computed before[4] and are negative in accord with the repulsive nature of Hartree potential. The next column, $d=1.225$, has a special significance for the positron-hydrogen problem. The fact that these phase shifts are closer to and larger than Rotenberg's may be of significance.

The argument from which the value $d\approx1.225$ is derived is surely very old.[5] It is based on the observation that the potential which the positron sees (exclusive of its interaction with the nucleus) is identical to the energy of the electron in the combined field of the positron and the nucleus. However this energy is trivially calculable in the limit that the positron approaches the nucleus. In that case the electron sees simply a coulomb field of charge $Z=2$. In other words we must have

$$\lim_{r\to 0}\left\{\left[2e^{-2r}\left(1+r^{-1}\right)-2r^{-1}\right]-\frac{\alpha}{\left(r^2+d^2\right)^2}\right\}=-\frac{Z^2}{n^2}.$$

where n is the principal quantum member of the electron in its ground state (if we are dealing with the scattering by hydrogen in its ground state). Therefore using the value $n=1$, $Z=2$, and $\alpha=\frac{9}{2}$, the known polarizability of hydrogen, we can solve for d: $d=(\frac{9}{4})^{1/4}\approx1.225$.

The remainder of this paper will be concerned with nonadiabatic theory of positron-hydrogen scattering. The quantitative results of this investigation are restricted to $k=0.2$, since we are mainly concerned with establishing whether Schwartz's or Rotenberg's phase shifts are more correct. The calculations herein support the latter's results and, to that extent, agree with the phenomenological result previously stated.

THE NONADIABATIC THEORY

The nonadiabatic theory starts with the decomposition of the s-wave function,*

$$\Psi(r_1, r_2, \theta_{12}) = \frac{1}{r_1 r_2}\sum_{l=0}^{\infty}(2l+1)^{1/2}\Phi_l(r_1 r_2)P_l(\cos\theta_{12}), \tag{1}$$

which is complete in view of the fact that the Hamiltonian

$$H = -\frac{1}{r_1}\frac{\partial^2}{\partial r_1^2}r_1 - \frac{1}{r_2}\frac{\partial^2}{\partial r_2^2}r_2 - \left(\frac{1}{r_1^2}+\frac{1}{r_2^2}\right)\frac{1}{\sin\theta_{12}}\frac{\partial}{\partial\theta_{12}}\sin\theta_{12}\frac{\partial}{\partial\theta_{12}} + \frac{2}{r_1} - \frac{2}{r_2} - \frac{2}{r_{12}}$$

depends only upon the three coordinates r_1, r_2, θ_{12}. As in electron-hydrogen problems these coordinates refer to the distance from the nucleus of the incoming particle (*positron* in this case) and the bound electron, and the angle between the respective radius vectors. One of the two salient differences between this and the electron-hydrogen problem is the changed signs of $2/r_1$ and $-2/r_{12}$ potentials. Here they are repulsive and attractive respectively. Thus the Hamiltonian is manifestly unsymmetrical with respect to r_1 and r_2, which implies that $\Psi(r_1, r_2, \theta_{12})$ cannot be symmetrical—the second difference. Both these differences have profound consequences on the present problem as compared with the electron-hydrogen problem.

Substitution of equation 1 into the Schrödinger equation

$$H\Psi = E\Psi \tag{2}$$

gives rise to the coupled set of equations

$$\left[\frac{\partial^2}{\partial r_1^2} + \frac{\partial^2}{\partial r_2^2} - l(l+1)\left(r_1^{-2}+r_2^{-2}\right) + E + \frac{2}{r_2} - \frac{2}{r_1} + M_{ll}\right]\Phi_l = -\sum_{m(\pm l)} M_{lm}\Phi_m , \tag{3}$$

where

$$M_{lm} = \sqrt{(2l+1)(2m+1)}\sum_{n=0}^{l+m}\frac{r_<^n}{r_>^{n+1}}C_{lmn} ,$$

with

$$C_{lmn} = \int_0^{\pi} P_l(\cos\theta)\,P_m(\cos\theta)\,P_n(\cos\theta)\sin\theta\,d\theta$$

the integral of three Legendre polynomials.[7] The dependence of the M_{lm} on $r_<^n/r_>^{n+1}$ means that there are two different equations depending on whether r_1 or r_2 is the larger. The equation for Φ_0, for example, is more explicitly:

$$\left(\Delta_{12} + E + \frac{2}{r_2}\right)\Phi_0^{>} = -\sum_{m=1}^{\infty} 2(2m+1)^{-1/2}\frac{r_2^m}{r_1^{m+1}}\Phi_m^{>} ,$$

$$\left(\Delta_{12} + E - \frac{2}{r_1} + \frac{4}{r_2}\right)\Phi_0^{<} = -\sum_{m=1}^{\infty} 2(2m+1)^{-1/2}\frac{r_1^m}{r_2^{m+1}}\Phi_m^{<} , \tag{4}$$

* See Appendix A for list of symbols.

where

$$\Phi_l = \begin{cases} \Phi_l^< & r_1 < r_2 \\ \Phi_l^> & r_1 > r_2 \end{cases} .$$

The boundary conditions are $\Phi_l(r_1 0) = 0 = \Phi_l(0, r_2)$ as long as we are below the threshold for inelastic scattering. Also, we have

$$\lim_{r_1 \to \infty} \Phi_l = \sin\left(kr_1 + \delta\right) R_{1s}(r_2)\, \delta_{l0}$$

and

$$\lim_{r_2 \to \infty} \Phi_l(r_1 r_2) = 0 .$$

The latter condition corresponds to the statement that below the inelastic threshold there are no states of a bound positron and a free electron; thus the wave function must vanish in the proper limit.

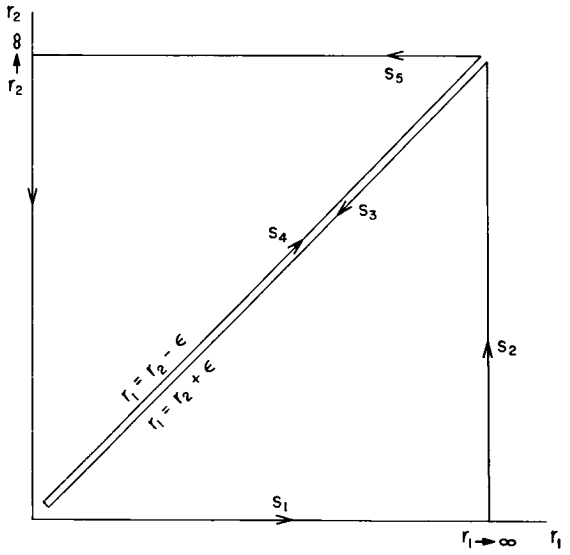


FIGURE 2.—The domains of integration in obtaining the basic relation (equation 7) for positron-hydrogen scattering.

Since we have, in effect, two equations in two continuous regions, appropriate matching conditions are required. Obviously they are

$$\Phi_l^< \Big|_{r_1 = r_2} = \Phi_l^> \Big|_{r_1 = r_2} ,$$

and

$$\left[\frac{\partial}{\partial n} \Phi_l^<\right]_{r_1 = r_2} - \left[\frac{\partial}{\partial n} \Phi_l^>\right]_{r_1 = r_2} . \tag{5}$$

The zeroth order problem is defined by neglecting the right side of equation 4:

$$\left(\Delta_{12} + E + \frac{2}{r_2}\right) \Phi_0^{(0)>} = 0, \qquad r_1 > r_2 \tag{6a}$$

$$\left(\Delta_{12} + E - \frac{2}{r_1} + \frac{4}{r_2}\right) \Phi_0^{(0)<} = 0. \qquad r_1 < r_2 \tag{6b}$$

The boundary conditions for $\Phi_0^{(0)}$ are in complete analogy to those for Φ_0. In particular the zeroth order phase shift δ_0 comes from the requirement that

$$\lim_{r_1 \to \infty} \Phi_0^{(0)} = \sin\left(kr_1 + \delta_0\right) R_{1s}\left(r_2\right) .$$

The zeroth order problem corresponds to a hydrogen atom and an incoming positron which sees no charge at all as long as it is outside the orbital electron, but which sees a repulsive (nuclear) potential when it is inside the electron. For this configuration the electron sees a doubly charged nucleus.

The basic relation between δ and δ_0 is obtained by the usual procedure of premultiplication and integration of the $\Phi_0^{(0)}$ equation by Φ_0 and vice-versa (see reference 6). In this case the domain integration is shown in figure 2. The area integrals involving Δ_{12} are converted to line integrals, and we use the diagonal boundary conditions (equation 5) for Φ_0 and the counterpart for $\Phi_0^{(0)}$,

$$\Phi_0^{(0)<}\Big|_{r_1=r_2} = \Phi_0^{(0)>}\Big|_{r_1=r_2} ,$$

$$\left[\frac{\partial}{\partial n}\Phi_0^{(0)<}\right]_{r_1=r_2} = \left[\frac{\partial}{\partial n}\Phi_0^{(0)>}\right]_{r_1=r_2} ,$$

to show that the line integrals along s_3 and s_4 cancel each other. The only contribution comes from s_2, and we obtain

$$\sin\left(\delta - \delta_0\right) = k^{-1} \sum_{m=1}^{\infty} 2(2m+1)^{-1/2} \int_0^{\infty}\int_0^{\infty} dr_1\, dr_2\, \Phi_0^{(0)} \frac{r_<^m}{r_>^{m+1}} \Phi_m \quad . \tag{7}$$

CALCULATIONS INVOLVING THE NONADIABATIC THEORY

From the verbal intepretation of the zeroth order problem we see clearly that it is a much cruder approximation of the complete positron-hydrogen scattering problem than the corresponding zeroth order problem for electron-hydrogen scattering. In particular it neglects the attraction of the positron

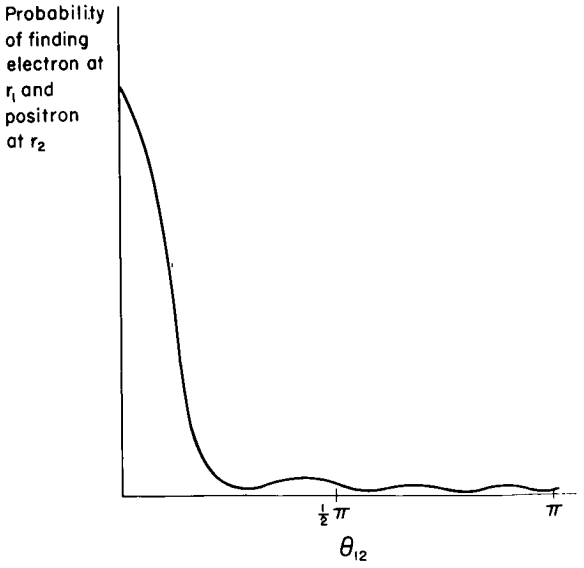


FIGURE 3.—The probability of finding a positron and a hydrogenic electron at distances r_1 and r_2 from the origin as a function of the angle between the radii and vectors.

to the electron and thus gives rise to negative phase shifts (for δ_0), whereas the work of Spruch and Rosenberg[1] shows that the final δ must be positive as k approaches 0. The inadequacy in the description of the region $r_1 \approx r_2$ has a further consequence which can be demonstrated pictorially.

In figure 3 we have sketched the probability of finding the electron at r_1 and the positron at r_2 as a function of θ_{12}. As a result of the electron-positron attraction this probability must rise sharply for $\theta_{12} \approx 0$. Remembering that the basic decomposition (equation 1) is in terms of Legendre polynomials, imagine that figure 1 refers to an angular distribution curve for a differential scattering process. Such a process would be described as having large forward angle scattering, and this in turn would mean the contribution of many partial waves. In terms of the present formalism this corresponds to the contribution of many relative partial waves, and this means that equation 2 cannot be expected to converge as rapidly as in the electron–hydrogen problem.

In the electron–hydrogen calculation, we also assumed, in addition to the assumption that there is a reasonably uniform convergence in Φ_l for increasing l, that the contribution from the configuration space $r_2 \approx r_1$ was small. The above discussion shows that the latter approximation is not justified. We shall therefore confine ourselves to solving the coupled set of equations;

$$\left(\Delta_{12} + E + \frac{2}{r_2} - \frac{2}{r_1} + \frac{2}{r_>}\right)\Phi_0 = -\frac{2}{\sqrt{3}}\frac{r_<}{r_>^2}\Phi_1 , \tag{8a}$$

$$\left(\Delta_{12} + E + 2\left(\frac{1}{r_1^2} + \frac{1}{r_2^2}\right)\right.$$

$$\left. + \frac{2}{r_2} - \frac{2}{r_1} + \frac{2}{r_>} + \frac{4}{5}\frac{r_<^2}{r_>^3}\right)\Phi_1 = -\frac{2}{\sqrt{3}}\frac{r_<}{r_>^2}\Phi_0 , \tag{8b}$$

which were obtained by truncating equation 3 beyond Φ_1 but without making any assumption about $r_<^n / r_>^{n+1}$.

First it is necessary to solve the zeroth order problem defined by equation 6 and the boundary conditions. Assuming that $\Phi_0^{(0)}$ is known, we can consider a first approximation of Φ_1 as the solution of

$$\left(\Delta_{12} - 2\left(\frac{1}{r_1^2} + \frac{1}{r_2^2}\right) + E + \frac{2}{r_2} - \frac{2}{r_1} + \frac{2}{r_>} + .8\frac{r_<^2}{r_>^3}\right)\Phi_1^{(0)} = \frac{-2}{\sqrt{3}}\frac{r_<}{r_>^2}\Phi_0^{(0)} . \tag{9}$$

The following sum rules are useful in the subsequent development:

$$\int_0^\infty\int \Phi_0^{(0)}\left[-2\left(\frac{1}{r_1^2} + \frac{1}{r_2^2}\right) + .8\frac{r_<^2}{r_>^3}\right]\Phi_1^{(0)}\, dr_1\, dr_2 = \frac{2}{\sqrt{3}}\int_0^\infty\int \frac{r_<}{r_>^2}\left(\Phi_0^{(0)}\right)^2 dr_1\, dr_2 , \tag{10a}$$

$$\int_0^\infty\int \Phi_0\left[-2\left(\frac{1}{r_1^2} + \frac{1}{r_2^2}\right) + .8\frac{r_<^2}{r_>^3}\right]\Phi_1\, dr_1\, dr_2 = \frac{2}{\sqrt{3}}\int_0^\infty\int \frac{r_<}{r_>^2}\left[(\Phi_0)^2 - (\Phi_1)^2\right] dr_1\, dr_2 . \tag{10b}$$

From equations 8b and 9 we can also derive

$$\int_0^\infty\int \left(\Phi_1^{(0)}\Phi_0 - \Phi_1\Phi_0^{(0)}\right)\frac{r_<}{r_>^2}\, dr_1\, dr_2 = 0 . \tag{11}$$

We can write

$$\Phi_1 = \Phi_1^{(0)} + \Delta\Phi_1 ,$$

$$\Phi_0 = \Phi_0^{(0)} + \Delta\Phi_0 ;$$

Equation 11 now implies

$$\int_0^\infty \int \frac{r_<}{r_>^2} \Phi_1^{(0)} \Delta\Phi_0 \, dr_1 \, dr_2 = \int_0^\infty \int \frac{r_<}{r_>^2} \Phi_0^{(0)} \Delta\Phi_1 \, dr_1 \, dr_2 . \tag{12}$$

A sufficient condition for the satisfaction of equation 12 is

$$\Phi_1^{(0)} \Delta\Phi_0 = \Phi_0^{(0)} \Delta\Phi_1 .$$

This equality is almost certainly too strong to be exactly true, but it is difficult to see why it should not be satisfactory as an approximation. Further, writing

$$\Delta\Phi_1 = C \Phi_1^{(0)} g(r_1, r_2) , \tag{13}$$

we can put equation 10b into the form:

$$2C \int_0^\infty \int \Phi_0^{(0)} \left[2\left(\frac{1}{r_1^2} + \frac{1}{r_2}\right) - .8 \frac{r_<^2}{r_>^3} \right] \Phi_1^{(0)} g(r_1 r_2) \, dr_1 \, dr_2$$

$$+ C^2 \int_0^\infty \int \Phi_0^{(0)} \left[2\left(\frac{1}{r_1^2} + \frac{1}{r_2^2}\right) - .8 \frac{r_<^2}{r_>^3} \right] \Phi_1^{(0)} \left[g(r_1 r_2) \right]^2 dr_1 \, dr_2$$

$$= -\frac{2}{\sqrt{3}} \int_0^\infty \int \frac{r_<}{r_>^2} (\Phi_1^{(0)})^2 \, dr_1 \, dr_2 + \frac{4C}{\sqrt{3}} \int_0^\infty \int \frac{r_<}{r_>^2} \left[(\Phi_0^{(0)})^2 - (\Phi_1^{(0)})^2 \right] g(r_1 r_2) \, dr_1 \, dr_2$$

$$+ 2(3)^{-1/2} C^2 \iint \frac{r_<}{r_>^2} \left[(\Phi_0^{(0)})^2 - (\Phi_1^{(0)})^2 \right] \left[g(r_1 r_2) \right]^2 dr_1 \, dr_2 . \tag{14}$$

In writing $\Delta\Phi$ in the form of equation 13 we consider $g(r_1 r_2)$ a fixed function, and C a constant determined by the solution of the quadratic equation (equation 14). We have chosen

$$g(r_1 r_2) = \left(\frac{r_2}{r_1}\right) \exp\left(-\frac{r_2}{r_1}\right) . \tag{15}$$

This function vanishes at $r_1 = 0$, $r_2 = 0$; is continuous with continuous derivative across $r_1 = r_2$; and has a maximum at $r_1 = r_2$.

In order to evaluate the integrals in equation 14, it is also necessary to know $\Phi_1^{(0)}$. This can be handled just as in the electron-hydrogen problem,[6] viz., we append a parameter D to the adiabatic form of $\Phi_1^{(0)}$:

$$\Phi_1^{(0)} = \frac{2}{\sqrt{3}} \frac{\sin(kr_1 + \delta_0)}{r_1^2 + D} e^{-r_2\left(\frac{r_2^3}{2} + r_2^2\right)} ,$$

where D is determined such that equation 8a is satisfied. Thus we are left with the difficult problem of evaluating $\Phi_0^{(0)}$. The expansion used in the electron-hydrogen case simply will not work here. In

the region $r_1 < r_2$ there is only one separable solution of equation 6b; in addition, it is difficult to formulate a useful expression which will minimize the deviation from both boundary conditions simultaneously (equation 5).

It should be noted, however, that as in the electron-hydrogen problem[6] any attack on the complete problem with functions of r_1 and r_2 can only approximate the zeroth order problem. At least three approximations have been carried out which fall into this category. One is the static approximation, the second an (unsymmetrical) close-coupling approximation with $1s$–$2s$–$3s$ hydrogenic states included,[4] and the third that part of Rotenberg's calculation[3] in which no θ_{12} dependence is contained. It was originally thought that the static approximation would be sufficient for quantitative purposes. The supposition seemed to be confirmed by the fact that the inclusion of the $2s$ and $3s$ states[4] increased the static phase shifts by less than 4 percent.

Therefore, in the first instance we used the $1s$–$2s$–$3s$ approximation for $\Phi_0^{(0)}$:

$$\Phi_0^{(0)} = \sum_{n=1}^{3} u_n(r_1) R_{ns}(r_2) . \tag{16}$$

The integrals in the above development could then be evaluated. In order to satisfy the dipole sum rule (equation 10a) we found it necessary to augment $\Phi_1^{(0)}$. Let $\tilde{\delta}$ be the phase shift of the problem defined by the coupled set (equation 8). The, according to the above development,

$$\sin(\tilde{\delta} - \delta_0) = \frac{2}{k\sqrt{3}} \int_0^\infty \int \Phi_0^{(0)} \frac{r_<}{r_>^2} \Phi_1^{(0)} dr_1 dr_2 + \frac{2}{k\sqrt{3}} \int_0^\infty \int \Phi_0^{(0)} \frac{r_2}{r_>^2} C\Phi_1^{(0)} g\, dr_1 dr_2 . \tag{17}$$

At $k=0.2$ we obtained the values 0.1414 and 0.0042 for the two terms on the right side of equation 17. With $\delta_0 = -0.1102$ as given by the above $\Phi_0^{(0)}$, we obtained

$$\tilde{\delta} = \delta_0 + \sin^{-1}(0.1456) = 0.0359 .$$

The value of δ obtained by Spruch and Rosenberg[1] at $k=0.2$ is 0.150. This value can be taken as a reliable (but not rigorous) lower bound on the phase shift. As mentioned previously, the multipole expansion cannot be expected to converge as dramatically as in the electron-hydrogen case; nevertheless the difference between these two numbers seemed greater than could be attributed to higher multipoles.

To check this, we computed several terms of equation 7 using the appropriate projections of the Spruch–Rosenberg wave function[1] to obtain Φ_l and using equation 16 for $\Phi_0^{(0)}$. The results for $k=0.2$ are given in table 2. The last row is obtained from equation 7 using the first three terms and the appropriate value of δ_0 (in the $l=0$ row). This δ is reasonably close to 0.150 which Spruch and Rotenberg obtained from their complete calculation. The results indicate that the contribution of the higher multipoles is fairly small—although not negligible—and that the main discrepancy is in the dipole contribution: 0.194 versus our 0.146. There could be many causes for this discrepancy and for the apparent consistency of the Spruch–Rosenberg wave function with the Smith–Burke $\Phi_0^{(0)}$. As far as the present calculation was concerned, it was felt after much delibration that the approximation (equation 16) for $\Phi_0^{(0)}$ must somehow be inadequate.

This supposition has very recently been markedly supported by the work of Rotenberg.[3] Insofar as it bears on the zeroth order problem, his

TABLE 2.—*Results of Spruch–Rosenberg Wave Function for $k=0.2$*

l (radians)	Multipole Integral
0	−0.1102
1	0.194
2	0.0377
3	0.00947
δ	0.131

TABLE 3.—*Effect of Adding Terms*

Type of Function	Total Number of Terms*							
	1	2	3	4	5	6	7	8
Sturmian (row b)	−0.067	−0.041	−0.022	0.057	0.093	0.195	0.212	0.232
Hydrogenic (row c)	−0.1145	−0.1109	−0.1102	-	-	-	-	-

*Taken successively as (n, L) = (1, 0), (2, 0), (3, 0), (2, 1), (3, 1), (4, 1), (3, 2), (4, 2).

expansion is of the form of equation 16 with the hydrogenic functions $R_{ns}(r_2)$ replaced by the Sturmian functions $S_{n0}(r_2)$. In this approximation

$$\Phi_0^{(0)} = \sum_{n=1}^{3} u_n(r_1)\, S_{n0}(r_2) . \qquad (18)$$

$S_{10}(r_2)$ is identical to R_{1s}, and the remaining sturmian functions (with our normalization) are:

$$S_{20}(r) = 2\sqrt{2}\, e^{-r}(r - r^2) ,$$

$$S_{30}(r) = 2(3)^{-1/2} e^{-r}(3r - 6r^2 + 2r^3) .$$

The chief property of the Sturmian function is that they are a complete set without a continuous spectrum.[3] The effect of this expansion on the zeroth order phase shift can be gleaned from the first three columns of row b, table 3 (taken from table II of reference 3) for $k=0.2$. The first three columns refer to the zeroth order problem, and are to be compared to successive augmentation of the hydrogenic wave functions $1s$, $2s$, $3s$ in the close coupling expansion row c.[4] Then, according to Rotenberg's results, the convergence of the close-coupling expansion is spuriours, and the correct zeroth order phase shift is much less negative than had hitherto been thought.* Using equation 18 for $\Phi_0^{(0)}$, we obtain from the nonadiabatic theory, as outlined above, the contributions 0.227 and 0.0169 for the two terms of equation 17. This in turn yields a $\tilde{\delta}=0.224$.

The calculation was repeated using the Ansatz

$$\Phi_1^{(0)} = \frac{2}{\sqrt{3}} \frac{u_1(r_1)}{r_1^2 + D} e^{-r_2} \left(\frac{r_2^3}{2} + r_2^2 \right) ,$$

where $u_1(r_1)$ is first function in equation 18. In this case we obtained the values 0.222 and 0.0172 and a final $\tilde{\delta}=0.219$. These results agree to two figures and are compared with the entry in column 6 ($\tilde{\delta}=0.195$) of table 3. Up to this point only relative s- and p-states have been included; so that, through column 6, Rotenberg approximates the solution of the coupled set (equations 8). The comparison shows that with the inclusion of three (n, l) terms he has almost obtained convergence, relative to having Φ_0 described by three $(n, 0)$ terms. The subsequent addition of d-wave terms gives an additional contribution 0.037 to this phase shift. This agrees identically to the $l=2$ contribution (table 2) of the Spruch–Rosenberg wave function. The exact agreement is undoubtedly coincidental, but it adds confidence that this contribution, too, is well represented by the Sturmian expansion.

It follows that, if the Sturmian expansion (equation 18) for $\Phi_0^{(0)}$ is essentially correct, then Rotenberg's phase shifts are more correct than those of Schwartz (reference 2; see also table 1) and that the former are very likely a lower bound.

Two questions immediately present themselves. First, how adequate is the Sturmian expansion of $\Phi_0^{(0)}$? The lack of convergence of Rotenberg's results is most noticeable in this zeroth order

* Nevertheless, as long as the phase shifts are negative, we can find a physical rationalization for the result; the part of the positron's wave function inside the orbital electron makes the latter see effectively a more positively charged nucleus. This makes the electron more tightly bound; thus the part of the positron's wave function outside of the electron sees a repulsive but much smaller structure, giving rise to negative but smaller (in absolute value) phase shifts. It is evident that such a picture corresponds to a highly nonadiabatic situation.

results. Indeed, naively extrapolating the first three columns of table 3 might lead to a positive zeroth order phase shift—a conclusion which, in view of the physical interpretation of the zeroth order problem, is almost inconceivable. We suspect that the convergence of the Sturmian expansion is exceedingly rapid, so that within the addition of the first couple of terms the phase shift rapidly approaches the converged value and that the change, after a couple of terms have been added, is very small. If this is the case, then we must sincerely applaud this remarkable expansion. For the present, however, the possibility cannot be excluded that the inclusion of more $(n, 0)$ Sturmian functions will make the phase shift retreat towards the close-coupling value.

The second question is, if Rotenberg's results[3] are basically correct, how can Schwartz's results (with many parameters) be so incorrect? This is a difficult question to answer. However, it should be noted that the number of terms in his wave function relating exclusively to the zeroth order problem is not overly great. Therefore his calculation may be inadequate precisely because it does not describe sufficiently the correlations of the zeroth order problem. If this is the case, it provides an additional reason for considering the zeroth order problem separately from the remainder of the scattering problem—which is the essence of the nonadiabatic theory.

In the recent work of reference 8, an adiabatic polarization potential $\alpha(r)/r^4$, where

$$\alpha(r) = \frac{9}{2} - e^{-2r}\left(\frac{2r^5}{3} + 3r^4 + 6r^3 + 9r^2 + 9r + \frac{9}{2}\right),$$

is used in place of the phenomenological potential $\alpha/(r^2+d^2)^2$. The latter was derived[9]* specifically for the problem of *electron-hydrogen* scattering[10] as a prototype of *electron-atom* scattering.[11] In that case it was shown[6] that, beyond the spherically symmetric zeroth-order approximation, the phase shift is dominated by the dipole correction. Now the $\alpha(r)/r^4$ potential goes to zero at $r=0$ whereas the $\alpha/(r^2+d^2)^2$ potential does not (see figure 1). Thus the effect of the attraction in the former is suppressed for small r and phenomenologically simulates the contribution of only the dipole $(l=1)$ term in the nonadiabatic description. However, from table 2 it is clear that in positron-hydrogen scattering a significant contribution comes from the higher relative partial waves. The stronger dependence of the $\alpha/(r^2+d^2)^2$ potential near the origin appears better suited to simulate this contribution.

NONADAIABATIC THEORY OF HIGHER PARTIAL WAVE SCATTERING

Let Ψ_L be the wave function of the partial wave angular momentum L. For the elastic scattering problem,

$$\lim_{r_1\to\infty} \Psi_L = \frac{\sin\left(kr_1 - \frac{1}{2}\pi l + \delta^{(L)}\right)}{r_1} Y_{L_0}(\Omega_1)\, \frac{R_{1s}(r_2)}{r_2} Y_{00}(\Omega_2) .$$

We can expand this wave function

$$\Psi_L = \frac{1}{r_1 r_2} \sum_{l_1, l_2} \Phi_{l_1 l_2}^{(L)}(r_1 r_2) \sum_{m_1} \left(l_1 l_2 m_1 - m_1 \mid L0\right) Y_{l_1 m_1}(\Omega_1) Y_{l_2 -m_1}(\Omega_2) , \qquad (19)$$

where $(l_1 l_2 m_1 m_2 \mid LM)$ are the Clebsh–Gordon coefficients in the notation of Condon and Shortley.[7]

* The potential $\alpha(r)/r^4$ was derived by multiplying the adiabatic dipole term in the wave function by a step function. This function has a discontinuity in value and an infinite discontinuity in slope at $r_1=r_2$. Both these features are removed in the e^-–H problem by (anti) symmetrizing the total wave function. However, in the e^+–H problem, the ansatz for the wave function retains these unpleasant features, since there is no symmetrization.

With this expansion the Schrödinger equation can be reduced to

$$\left[\Delta_{12} - \frac{l_1(l_1+1)}{r_1^2} - \frac{l_2(l_2+1)}{r_2^2} - \frac{2}{r_1} + \frac{2}{r_2} + E + \left\langle l_1\, l_2 \left| \frac{2}{r_{12}} \right| l_1\, l_2 \right\rangle \right] \Phi_{l_1 l_2}^{(L)} = - \sum_{\lambda_1, \lambda_2}{}' \left\langle l_1\, l_2 \left| \frac{2}{r_{12}} \right| \lambda_1 \lambda_2 \right\rangle \Phi_{\lambda_1 \lambda_2}^{(L)} , \quad (20)$$

where

$$\left\langle l_1\, l_2 \left| \frac{2}{r_{12}} \right| \lambda_1 \lambda_2 \right\rangle \equiv \sum_{m_1} \left(l_1\, l_2\, m_1 - m_1 \middle| L0 \right)^2 \iint Y^*_{l_1 m_1}(\Omega_1)\, Y^*_{l_2 - m_1}(\Omega_2)\, \frac{2}{r_{12}}\, Y_{\lambda_1 m_1}(\Omega_1)\, Y_{\lambda_2 - m_1}(\Omega_2)\, d\Omega_1\, d\Omega_2 .$$

The quantities l_1, l_2, and L must satisfy the usual triangular equality. In addition l_1+l_2 and L must have the same parity. From equation 20 it is clear that the effective centrifugal barrier which a given $\Phi_{l_1 l_2}{}^{(L)}$ sees is determined by l_1+l_2. For a given L ($L \neq 0$) there are two functions which are associated with the lowest centifugal barrier ($l_1+l_2=L$), $\Phi_{L0}{}^{(L)}$ and $\Phi_{0L}{}^{(L)}$. The rule of parity then says that the next functions which enter the expansion (equation 19) must have a centrifugal barrier *two* units greater than L. Thus we have a powerful argument that the higher partial wave scattering is determined mainly by the coupled set of equations for the lowest two functions.

Applying this approximation to p-wave scattering (and suppressing the superscript 1), we get:

$$\left(\Delta_{12} - \frac{2}{r_1^2} - \frac{2}{r_1} + \frac{2}{r_2} + \frac{2}{r_>} + E\right) \Phi_{10} = -\left(\frac{2}{3}\right)\left(\frac{r_<}{r_>^2}\right) \Phi_{01} ,$$

$$\left(\Delta_{12} - \frac{2}{r_2^2} - \frac{2}{r_1} + \frac{2}{r_2} + \frac{2}{r_>} + E\right) \Phi_{01} = -\left(\frac{2}{3}\right)\left(\frac{r_<}{r_>^2}\right) \Phi_{10} \quad (21)$$

where the boundary conditions are

$$\lim_{r_1 \to \infty} \Phi_{10} = \sin\left(kr_1 + \delta^{(L)} - \frac{\pi}{2}\right) R_{1s}(r_2) ,$$

$$\lim_{r_1 \to \infty} \Phi_{01} = 0 , \quad (22)$$

plus the vanishing of these functions in all other asymptotic regions.

Now since both Φ_{10} and Φ_{01} are associated with the same amount of centifugal barrier, we expect a reliable solution will be reasonably simultaneous. As a very first approximation, however, we may note that there is a great asymmetry in the boundary conditions, equation 22, and between Φ_{10} and Φ_{01}, leading to the qualitative expectation that the former function is dominant over the latter. Within this approximation we could consider a zeroth order problem:

$$\left(\Delta_{12} - \frac{2}{r_1^2} - \frac{2}{r_1} + \frac{2}{r_2} + \frac{2}{r_>} + E\right) \overset{0}{\Phi}_{10} = 0 , \quad (23)$$

with the asymptotic solution

$$\lim_{r_1 \to \infty} \overset{0}{\Phi}_{10} = \sin\left(kr_1 - \frac{\pi}{2} + \delta_0^{(1)}\right) R_{1s}(r_2) .$$

We can then readily derive

$$\sin\left(\tilde{\delta}^{(1)} - \delta_0^{(1)}\right) = \frac{2}{3k} \int_0^\infty \int \overset{0}{\Phi}_{10} \frac{r_<}{r_>^2} \Phi_{01} \, dr_1 \, dr_2 . \tag{24}$$

from equations 21 and 23. Equation 24 is exact; the approximation comes in when we replace Φ_{01} by the approximation $\overset{0}{\Phi}_{01}$, which is governed by

$$\left(\Delta_{12} - \frac{2}{r_2^2} - \frac{2}{r_1} + \frac{2}{r_2} + \frac{2}{r_>} + E\right) \overset{0}{\Phi}_{01} = -\frac{2}{3} \frac{r_<}{r_>^2} \overset{0}{\Phi}_{10}$$

From this equation and equation 23, we obtain

$$\int_0^\infty \int \overset{0}{\Phi}_{10} \left(\frac{2}{r_2^2} - \frac{2}{r_1^2}\right) \overset{0}{\Phi}_{01} \, dr_2 \, dr_2 = \frac{2}{3} \int_0^\infty \int \left(\overset{0}{\Phi}_{10}\right)^2 \frac{r_<}{r_>^2} \, dr_1 \, dr_2 . \tag{25}$$

The following approximations are used in solving for $\tilde{\delta}^{(1)}$:

$$\overset{0}{\Phi}_{10} = kr_1 \, j_1(kr_1) ,$$

where $j_1(kr)$ is the spherical Bessel function and, by implication, $\delta_0^{(1)} = 0$; and

$$\overset{0}{\Phi}_{01} = \frac{2}{3} \frac{kr_1 \, j_1(kr_1)}{r_1^2 + D} e^{-r_2} \left(\frac{r_2^3}{2} + r_2^2\right) ,$$

which is the adiabatic form of Φ_{01}. In principal D is to be determined by the satisfaction of equation 25. Actually we were unable to satisfy the sum rule with $\overset{0}{\Phi}_{01}$ in the above form. Therefore we had to change $\overset{0}{\Phi}_{01}$ to $\tilde{\Phi}_{01}$:

$$\tilde{\Phi}_{01} = \overset{0}{\Phi}_{01} \left(1 + Cg(r_1 r_2)\right) ,$$

where $g(r_1 r_2)$ is given in equation 15, D was chosen as unity, and C was varied to give equality. We can now determine $\tilde{\delta}^{(1)}$ by using

$$\tilde{\delta}^{(1)} \approx \frac{2}{3k} \int_0^\infty \int \overset{0}{\Phi}_{10} \frac{r_<}{r_>^2} \tilde{\Phi}_{01} \, dr_1 \, dr_2 .$$

In obtaining these results we have thus made two approximations. The first is the replacement of the exact p-wave equation by the truncated set (equation 21), which is equivalent to neglecting the difference between $\delta^{(1)}$ and $\tilde{\delta}^{(1)}$. We would expect that this would not introduce an error of more than 25 percent—erring on small side. We have also approximated the solution of equation 21. However, in this case $\delta_0^{(1)}$ is an over-estimate of the exact $\delta_0^{(1)}$, which must surely be negative. On the whole therefore the final phase shifts in table 4 can reasonably be expected to be within a factor 2 of the exact p-wave phase shifts. The discrepancy with Bransden's phase shifts (reference 12, and also table 4) is generally greater than this amount.

The phase shifts are essentially negligible as

TABLE 4.—*p-Wave Phase Shifts*

k (ryd$^{1/2}$)	Phase Shift (radians)	
	Nonadiabatic	Reference 12
0.1	0.00219	
0.2	0.00931	0.016
0.3	0.0181	
0.4	0.0277	0.104
0.5	0.0371	
0.6	0.04565	0.167
0.7	0.0531	

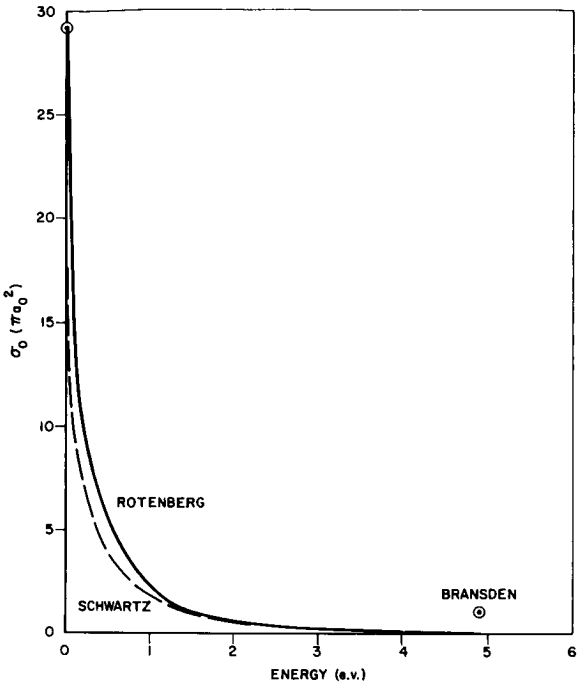


FIGURE 4.—The s-wave scattering cross section of positrons by hydrogen in two approximations. The nonadiabatic p-wave contribution is negligible on this scale whereas Bransden's p-wave (indicated by the point at 5 ev) is not.

far as the total cross section is concerned. The s-wave cross sections are presented in figure 4. It can be seen that it will be very difficult to distinguish experimentally between the results of Schwartz and Rotenberg. (There is a remote possibility at 0.5 ev.) On the other hand, in the region of the minimum (5 ev) the p-wave will yield a cross section of less than 0.5 πa_0^2 whereas, according to Bransden, the cross sectihn should be about πa_0^2. This second case should be experimentally distinguishable.

ACKNOWLEDGMENTS

We should like to thank Dr. Kenneth Smith and Dr. Manual Rotenberg for the numerical values of their respective $u_n(r_1)$ functions in the expansion of equations 16 and 18.

REFERENCES

1. SPRUCH, L., and ROSENBERG, L., "Low-Energy Scattering by a Compound System: Positrons on Atomic Hydrogen," *Phys. Rev.* **117**(1): 143–151, January 1, 1960.
2. SCHWARTZ, C., "Electron Scattering from Hydrogen," *Phys. Rev.* **124**(5): 1468–1471, December 1, 1961.
3. ROTENBERG, M., "Application of Sturmian Functions to the Schröedinger Three-Body Problem: Elastic e–H Scattering," *Ann. Phys.* **19**(2): 262–278, 1962.
4. SMITH, K., and BURKE, P. G., "Effect of Virtual Excitations on the Elastic Scattering of Electrons and Positrons by Atomic Hydrogen," *Phys. Rev.* **123**(1): 174–178, July 1, 1961.
5. MITTLEMAN, M. H., "Scattering of Charged Particles by Neutral Atoms," *Phys. Rev.* **113**(1): 198–211, January 1, 1959.
6. TEMKIN, A., "Nonadiabatic Theory of Electron-Hydrogen Scattering," *Phys. Rev.* **126**(1): 130–142, April 1, 1962.
7. CONDON, E. U., and SHORTLEY, G. H., "The Theory of Atomic Spectra," Cambridge: University Press, 1951.
8. CODY, W. J., and LAWSON, J., "The Elastic Scattering of Protons by Atomic Hydrogen," to be published, 1962.
9. TEMKIN, A., "A Note on the Scattering of Electrons from Atomic Hydrogen," *Phys. Rev.* **116**(2): 358–363, October 15, 1959.
10. TEMKIN, A., and LAMKIN, J. C., "Application of the Method of Polarized Orbitals to the Scattering of Electrons from Hydrogen," *Phys. Rev.* **121**(3): 788–794, February 1, 1961.
11. TEMKIN, A., "Polarization and Exchange Effects in the Scattering of Electrons from Atoms with Application to Oxygen," *Phys. Rev.* **107**(4): 1004–1012, August 15, 1957.
12. BRANSDEN, B. H., "The Low Energy Scattering of Positrons by Hydrogen Atoms," *Proc. Phys. Soc.* **79**(507): 190–195, January 1962.

Appendix A

LIST OF SYMBOLS

$P_l(\cos\theta_{12})$ Legendre polynomials of order l of $\cos\theta_{12}$.

r_1 The distance of positron from the nucleus.

r_2 The distance of electron from the nucleus.

$R_{ns}(r_2)$ The s-eigenstates of hydrogen; in particular $R_{1s}(r_2) = 2r_2^{-r_2}$ is the ground state.

δ Exact s-wave phase shift.

δ_0 Phase shift of the zeroth order problem; that is,

$$\lim_{r_1 \to \infty} \Phi_0^{(0)}(r_1, r_2) = \sin(kr_1 + \delta_0) R_{1s}(r_2)$$

θ_{12} The angle between the lines connecting electron and positron to the nucleus.

$\Phi_l(r_1 r_2)$ Functions in the basic expansion of

$$\Psi(r_1, r_2, \theta_{12}) = \frac{1}{r_1 r_2} \sum_{l=0}^{\infty} (2l+1)^{1/2} \Phi_l(r_1, r_2) P_l(\cos\theta_{12})$$

$\Psi(r_1 r_2 \theta_{12})$ Solution of the s-wave scattering problem.

$S_{n0}(r_2)$ The Sturmian functions of order $n0$.

RELATIVE PARTIAL WAVE THEORY OF DIATOMIC MOLECULES

A. TEMKIN

Goddard Space Flight Center

The method of expanding the Hamiltonian in (relative) partial waves and introducing similarly expanded but truncated series as approximate solutions is utilized in the case of $(H_2)^+$ as a prototype of diatomic molecules. Thirteen states have been examined at four internuclear separations and results compared to exact results. In particular, oscillator strengths have been computed, and various properties of the approximate solutions are exhibited which make this method potentially very useful especially for some astrophysical applications. A simple modification of the theory can be made which ensures that the oscillator strengths satisfy the sum rules.

INTRODUCTION

The collisions of hydrogen and helium atoms and ions can be described as virtual states of the resulting molecular complexes. It is becoming increasingly probable that emissions and absorptions of these quasi-diatomic molecules is the mechanism of much of the ultraviolet opacity of the sun and hotter stars. In the case of the sun, it has been pointed out[1] that transitions from the $^3\Sigma_u^+$ to $\rightarrow{}^3\Sigma_g^+$ of H_2 have many times the absorption coefficient as that due to H^- at λ 2000. Thus, these transitions are the main cause of opacity at these shorter wavelengths, a fact which lends support to the hypothesis that quasimolecules of hydrogen and helium are responsible for the opacity in the ultraviolet of hotter stars.[2] Bates[2] has calculated that transitions of quasi $(H_2)^+$ are also important in the hot stars.

Stecher and Milligan[2] have singled out virtual transitions of the molecule helium hydride singly ionized, $(HeH)^+$, in view of the fact that He and $(H)^+$ have the largest densities in these hot stars. This molecule has two orbital electrons and as such it is tempting to attack this problem with methods which have recently been applied in atomic physics. We have in mind the method of relative partial waves which was introduced some time ago to explain an excited state of singly ionized lithium.[3] We have recently generalized this method to the scattering problem (electrons from atomic hydrogen).[4] The effectiveness of the method stems from the fact that it contains a very straightforward method of successive approximations, and very powerful arguments can be invoked for their rapid convergence. In addition, even in zeroth order the method gives enough quantitative accuracy for many applications.

In view of the complexity of the molecular problem, however, it has seemed worthwhile to apply the method to some simpler one-electron molecules $(H_2)^+$ and $(HeH)^{++}$ for which exact results are available.[5,6]

This paper reports on the results for $(H_2)^+$. After most of these calculations were completed, we were informed that similar calculations have been carried out by Cohen and Coulson[7,8] and that the idea has been reintroduced many times.[7] Cohen and Coulson have concentrated on one internuclear distance ($R_{AB}=2$ in Bohr radii), and on the convergence of the wavefunctions and energies as one goes to higher-order approximations. Because of our own rather more pragmatic attitude toward this method, in terms of the applications mentioned above, our present calculations are restricted to lower order, but

cover four internuclear separations ($R_{AB}=4$, 3, 2, 1) and more states. In particular, we have worked toward the calculation of oscillator strengths.

In Sec. II, we outline briefly the method and derive the main equations specifically for the hydrogen molecular ion. In Sec. III, we discuss the solution of the equations and present results, mostly in graphical form, in comparison with the exact results, where available. Although the solutions and results have many attractive properties, one shortcoming of the usual approach is that the sum rules for oscillator strengths are not automatically satisfied. In Sec. IV, we describe a modification of the approach which guarantees this feature. Actual numerical results show that the individual oscillator strengths are also improved. Section IV contains some concluding remarks—specifically, the necessity of going from zeroth to first order in the case of heteronuclear molecules to get comparable accuracy to zeroth order in the homonuclear molecules. Some remarks on two-electron diatomic molecules are also made.

NONADIABATIC THEORY

We consider the electronic states of $(H_2)^+$ within the confines of the Born–Opperheimer approximation. In that case, the Hamiltonian can be written (our units throughout are lengths in Bohr radii, energies in rydbergs):

$$H=-\nabla^2-2/r_A-2/r_B+2/R_{AB}. \quad (2.1)$$

We have explicitly included the repulsion $2/R_{AB}$ between the nuclei, so that our energies below are potential energy curves for the various states. The Hamiltonian can be expanded in the form

$$H\equiv-\nabla^2-\sum_{l(\text{even})} V_l P_l(\cos\theta)+2/R_{AB}, \quad (2.2a)$$

where

$$V_l=\begin{cases} r^l/(R_{AB/2})^{l+1} & r<\frac{1}{2}R_{ab} \\ (R_{AB/2})^l/r^{l+1} & r>\frac{1}{2}R_{ab}. \end{cases} \quad (2.2b)$$

The starting point of the nonadiabatic theory is an expansion in terms of spherical harmonics of the wave-function. For gerade and ungerade state we have

$$\Psi_M^{(g)}=\sum_{n(\text{even})}(\varphi_n^{(r)}/r)\,Y_{nM}(\Omega), \quad (2.3a)$$

$$\Psi_M^{(u)}=\sum_{n(\text{odd})}(\varphi_n^{(r)}/r)\,Y_{nM}(\Omega). \quad (2.3b)$$

There is a well known degeneracy with respect to $\pm M$. Substituting into Schrödinger equation, we can develop an infinite coupled set of differential equations for the $\varphi_n(r)$:

$$\left[\frac{d^2}{dr^2}-\frac{l(l+1)}{r^2}+E-\frac{2}{R_{AB}}\right]\varphi_l$$

$$+\left(\frac{4\pi}{2l+1}\right)^{1/2}\sum_{m(\text{even})}\sum_n V_m C_{lmn}{}^{MOM}\varphi_n=0. \quad (2.4)$$

The constants $C_{lmn}{}^{MOM}$ are integrals over three spherical harmonics which can be evaluated:

$$C_{lmn}{}^{MOM}=\int Y_{lM}{}^*(\Omega)\,Y_{m0}(\Omega)\,Y_{nM}(\Omega)\,d\Omega$$

$$=[(2l+1)(2m+1)/4\pi(2n+1)]^{1/2}$$

$$\times(lm00/n0)(lm-M0/n-M). \quad (2.5)$$

The latter two factors are Clebsch–Gordan coefficients. The coefficients are automatically zero if the parity of n is different from $l+m$.

The approximation consists of replacing the infinite set of equations by a finite set by truncating the terms after a given $l=l_{max}$. The convergence of this procedure stems from the presence of the centrifugal term $-l(l+1)/r^2$ which guarantees that as one goes to higher and higher l values, the amplitude of the corresponding radial terms gets more and more depressed in the region of interaction. This argument is the analog of the well-known argument for the convergence of partial waves in scattering problems.

The zeroth order approximation consists of retaining only the lowest radial term which enters a given expansion. For comparison with the approximation in Sec. IV, we write down explicitly the zeroth-order equations for the symmetry types

Σ_g and Σ_u:

Σ_g:

$$[(d^2/dr^2)+6/R_{AB}+E_0]\varphi_0^{(0)}(r)=0 \qquad r<\tfrac{1}{2}R_{AB},$$

$$[(d^2/dr^2)+4/r-2/R_{AB}+E_0]\varphi_0^{(0)}(r)=0 \quad r>\tfrac{1}{2}R_{AB}.$$

Σ_u:

$$\left(\frac{d^2}{dr^2}-\frac{2}{r^2}+\frac{6}{R_{AB}}+\frac{64}{5}\frac{r^2}{{R_{AB}}^3}+E_0\right)\varphi_1^{(0)}=0 \qquad r<\tfrac{1}{2}R_{AB},$$

$$\left(\frac{d_i}{dr^2}-\frac{2}{r^2}+\frac{4}{r}+\frac{2}{5}\frac{{R_{AB}}^2}{r^3}\frac{-2}{R_{AB}}+E_0\right)\varphi_1^{(0)}=0 \quad r>\tfrac{1}{2}R_{AB}.$$

We have appended the subscript (superscript) zero to $E(\varphi_l)$ to emphasize that the respective quantity is different in this approximation from what one would get in any higher order approximation.

The equations are seen to have a different form in the region $r<\frac{1}{2}R_{AB}$ and $r>\frac{1}{2}R_{AB}$. This is the means by which the singularities at the two nuclei are here approximated. It should also be noted that the zeroth-order problems have simple physical interpretations which also help clarify the nature of the approximations. For instance, in the Σ_g case, the orbital electron sees a uniform potential while inside the sphere swept out by the nuclei, and it sees a doubly charged nucleus at the origin when it is outside of that sphere. All this corresponds to replacing the nuclei by a uniformly charged spherical shell as is well-known from potential theory. The remaining zeroth-order problems also correspond to simple *but not identical* physical models of the interaction (cf. Sec. IV).

The different forms of the equation in the two regions of configuration space motivate the method of numerical solution. Specifically, the outer equations were integrated from a point $r\gg\frac{1}{2}R_{AB}$ into $r=\frac{1}{2}R_{AB}$. The inner equations were integrated from 0 to $\frac{1}{2}R_{AB}$. These integrations were done as a function of E and the logarithmic derivatives at the point $r=\frac{1}{2}R_{AB}$ of the respective functions were plotted. The equations were integrated numerically on the IBM 7090 computer of the Theoretical Division of the Goddard Space Flight Center. (The inner part of the Σ_g equation has a trivial analytic solution which was used.)

Figure 1 is a facsimile of such plot. The points of intersection correspond to the eigenvalues, and the states were further identified by observing the number of nodes of the associated radical functions. The intersections on the right of Fig. 1 correspond to states which we did not bother further to identify. The figure illustrates one of the main practical advantages of the present method: *the same calculation which gives the lowest state (of a given symmetry) can also be used to give the excited states as well.* Also, the different eigenfunctions of a given symmetry will automatically be orthogonal.

Among the intersections of the curves in Fig. 1, none corresponds to any state that we could legitimately identify as $\Sigma_g(3d)$. This is not surprising in view of the fact that the angular dependence of such a state would be primarily P_2 $(\cos\theta)$ which is completely neglected in the zeroth-order approximation. However, when we go to first order,

$$\Psi_{\Sigma_g}^{(1)} \propto \varphi_0^{(1)}(r)\,Y_{00}(\Omega)+\varphi_2^{(1)}(r)\,Y_{20}(\Omega),$$

we do have provision for such a function and, on carrying through this calculation, we do find this state in addition to improved energies of the $\Sigma_g(ns)$ states (cf. table II).

It should be noted that in any order the equa-

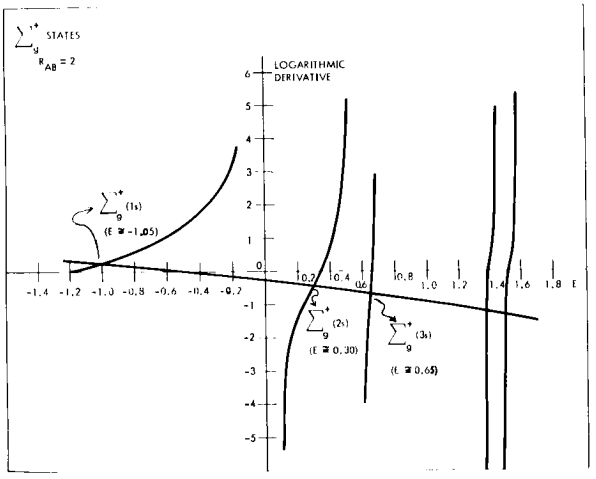


FIGURE 1.—Facsimile of inner (horizontal curve) and outer (vertical curves) logarithmic derivatives as functions of the energy E in rydbergs. Points of intersection correspond to bound states. The energies were ultimately determined more precisely than is here indicated.

tions that one derives are equivalent to a variational principle, so that the associated energies will always be upper bounds.

RESULTS

In Fig. 2, we give zeroth-order energies for the first three $\Sigma_g(ns)$ states (dashed curves) and exact energies (solid curves) as a function of R_{AB}. For the lowest $\Sigma_g(1s)$ state we also include the first-order and LCAO results. The first-order

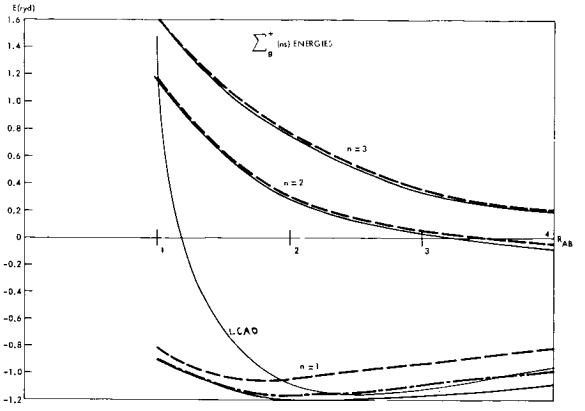


FIGURE 2.—Energies vs internuclear separation (R_{AB}) of the Σ_g^+ states. In this and the remaining figures, solid curves refer to exact results, dashed curves to zeroth order results, and dot-dash curves to first-order results. LCAO results are so labeled in the figures. Lengths are in Bohr radii.

results indicate the expected improvement in going to higher order. In table I, we have given numerical values of the energy of this state at $R_{AB}=2$ to third order, the higher-order results having been taken from the paper of Cohen and Coulson.[7] This inexorable convergence provides the theoretical backbone of the nonadiabatic approach. Table II contains the first-order energies of the $\Sigma_g(3d)$ state.

It should be emphasized, however, that in the limit of $R_{AB}\rightarrow\infty$, the nonadiabatic energy approaches zero[9] whereas in general it should correctly approach a negative number, which limit is correctly given by the LCAO result. In practice, therefore, a quite satisfactory fit of the energy should be obtainable by joining the zeroth order results at smaller R_{AB} to the LCAO results at larger R_{AB}.

TABLE I.—*The Convergence of the Nonadiabatic Approximations (Energy of the $\Sigma_g(1s)$ State, $R_{AB}=2$).*

Order	Zeroth	First	Second	Third[a]	Exact[b]
Energy	−1.038	−1.174		−1.199	−1.205

[a] Taken from Cohen and Coulson (reference 7).
[b] Bates, Ledsham, and Stewart (reference 5).

In Figs. 3–5, we have given the zeroth-order energies for the remaining symmetry classes that we have calculated. The exact excited state energies of Π_u and Π_g are not given by Bates et al.[5] Judging by the agreement of the higher states in the other cases, however, these values should be virtually exact.

One of the most important applications of the wavefunctions in astrophysics is the calculation of dipole integrals, Q, for the purpose of evaluating oscillator strengths. In the zeroth-order approximation this integral is (to within a proportionality constant)

$$Q \propto \int_0^\infty r\varphi_{l_i(\min)}{}^{(0)}\varphi_{l_f(\min)}{}^{(0)}\,dr. \qquad (3.1a)$$

Now, if we had used the complete wavefunction this expression would become

$$Q \propto \sum_{l_i}\sum_{l_f} a_i a_f \int_0^\infty r\varphi_{l_i}\varphi_{l_f}\,dr, \qquad (3.1b)$$

where a_i and a_f are constants. The zeroth-order approximation for Q thus involves *two* approximations. One is the neglect of all but the first term of Eq. (3.1b). The second is the use of the functions $\varphi_{l(\min)}{}^{(0)}$ in place of $\varphi_{l(\min)}$ in that

TABLE II.—*First-Order Energies[a] of the $\Sigma_g(3d)$ State*

R_{ab}	4	3	2	1
First order	−0.01814	0.1691	0.53105	1.54975
Exact[b]	−0.07145	0.1517	0.5285	1.54963

[a] These states do not exist in the zeroth-order approximation.
[b] Taken from reference 5.

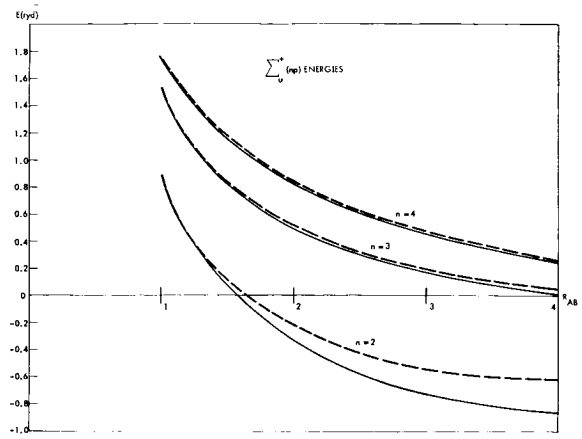


FIGURE 3.—$\Sigma_u(np)$ energies vs R_{AB}.

integral. The difference between those two functions stems from the fact that the latter includes all the back coupling of the higher φ_l on the first one. However, these functions are practically identical from the point of view of integration as demonstrated in Fig. 6. Here the exact $\varphi_0(r)$ [of the $\Sigma_g(2s)$ state] has been computed from the exact wavefunction $\Sigma_g(2s)$ of Bates et al.[5] by projection:

$$\varphi_0(r) \propto \int_0^{2\pi}\int_0^{\pi} \Psi_{\Sigma_g}(2s)\,(\mu, \lambda, \varphi)\,P_0\,(\cos\theta)\,\sin\theta\,d\theta\,d\varphi.$$

The connection between the prolate spheroidal coordinates μ, λ and the spherical coordinates r, θ is well known, and quadrature can be carried out numerically. The results show that the back coupling has very little effect on the form of the function.

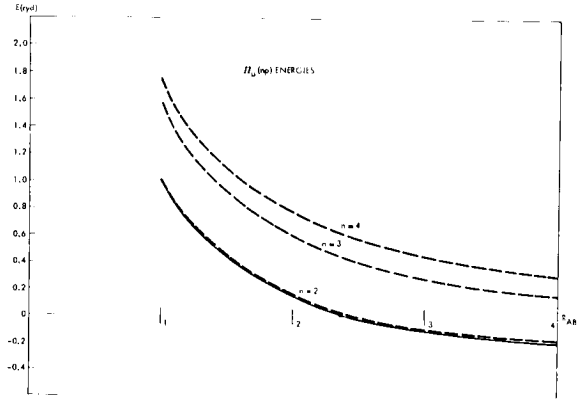


FIGURE 4.—$\Pi_u(np)$ energies vs R_{AB}.

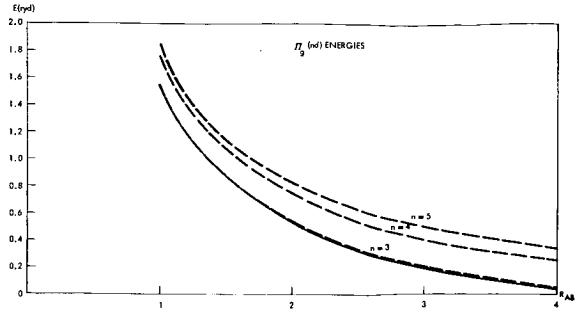


FIGURE 5.—$\Pi_g(nd)$ energies vs R_{AB}.

Thus the most significant approximation in using (3.1a) is the neglect of the higher terms which are present in (3.1b). The effect of this approximation can be seen in Fig. 7 in which Q vs R_{AB} is plotted for some of the dipole transitions we have calculated.

In Table III, we give the oscillator strengths from the ground, $\Sigma_g(1s)$, state. The partial sum that is included is the $n=2$, 3, 4, states of Σ_u and Π_u symmetry. If all states were included then to satisfy the Thomas–Kuhn sum rule, it would be necessary to have:

$$\sum f_{\Sigma_g-\Sigma_u} = \tfrac{1}{3}, \qquad (3.2a)$$

$$\sum f_{\Sigma_g-\Pi_u} = \tfrac{2}{3}, \qquad (3.2b)$$

where f refers to the individual oscillator strengths. In table III, it is to be observed that as R_{AB} increases, the individual sum rules are more poorly satisfied (which is to be expected) but that the sum of Σ_g–Σ_u and Σ_g–Π_u oscillator strengths gets closer to one.[10] In the next section, we show how the usual form of this approxima-

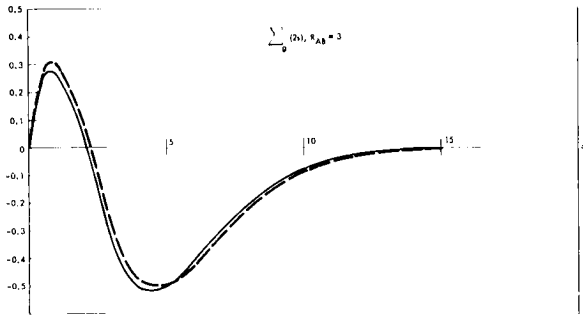


FIGURE 6.—Exact and zeroth-order radial functions vs the distance r of electron from center.

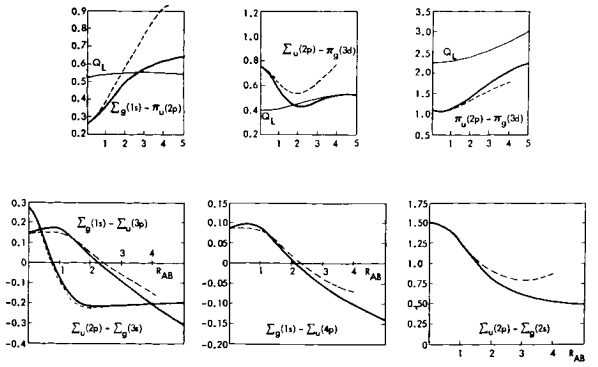


FIGURE 7.—Dipole integral Q vs R_{AB} for various transitions. Q_L refers to the LCAO results with the dipole length form of the matrix element. Q_L and the exact results are taken from reference (b) or (c) of Table III.

tion can be modified so that the satisfaction of the sum rules is guaranteed.

Table IV gives the remaining independent zeroth-order oscillator strengths in this approximation.

ALTERNATE APPROACH

The sum rules derived ultimately from the fact that one has a unified description of the interaction process for all states. We indicate in Sec. II that the zeroth-order problems do correspond to simple models of the interaction. The difficulty here is that each symmetry class corresponds to a different model. This defect can be remedied by truncating the Hamiltonian instead of the previous approach of truncating the expanded solutions. In zeroth-order, for example, all states would be governed by the Hamiltonian (2.2a) truncated after the first term $H_0 = -\nabla^2 - V_0 + 2/R_{AB}$. The associated Schrödinger equations, $H_0\Psi_0 = E_0\Psi_0$, are now solved exactly by our approach and each symmetry now corresponds to the *same* physical picture as previously described only the Σ_g states. The latter radial equation thus remains unaltered whereas the Σ_u equation (a tilde over the symmetry indicates this modified approximation) becomes:

$$[(d^2/dr^2) - (2/r^2) + (6/R_{AB}) + \tilde{E}_0]\tilde{\varphi}_1^{(0)} = 0 \qquad r < \tfrac{1}{2}R_{AB},$$

$$[(d^2/dr^2) - (2/r^2) + (4/r) - (2/R_{AB}) + \tilde{E}_0]\tilde{\varphi}_1^{(0)} = 0 \qquad r > \tfrac{1}{2}R_{AB}.$$

Comparison with the zeroth-order Σ_u equations shows that this equation lacks terms which are present there. One can also show that in this

TABLE III.—*Oscillator Strengths from the Ground State of* $(H_2)^+$

Transition	R_{AB}							
	4		3		2		1	
	Zeroth order	Exact	Zeroth order	Exact	Zeroth order	Exact	Zeroth order	Exact
$\Sigma_g(1s)$–Σ_u								
$2p^a$	0.120	0.238	0.174	0.289	0.230	0.319	0.239	0.269
$3p^b$	0.0063	0.016	0.002	0.0041	0.0007	0.00082	0.019	0.02
$4p^b$	0.0017	0.0046	0.0007	0.0015	0.00005	0.000055	0.005	0.006
$\Sigma_g(1s)$–Π_u								
$2p^c$	0.754	0.46	0.675	0.47	0.574	0.46	0.434	0.39
$3p^c$	0.053	...	0.0585	...	0.062	...	0.062	...
$4p^c$	0.014	...	0.017	...	0.019	...	0.022	...
Sum	0.95		0.93		0.89		0.78	

[a] Exact values from D. Bates, J. Chem. Phys. **19,** 1122 (1951).
[b] Exact values from D. Bates, R. Darling, S. Hawe, and A. Stewart, Proc. Phys. Soc. (London) **67,** 533 (1954).
[c] Exact values from D. Bates, R. Darling, S. Hawe, and A. Stewart, Proc. Phys. Soc. (London) **66,** 1124 (1953).

TABLE IV.—*Zeroth-Order Oscillator Strengths of* $(H_2)^+$

Transition	R_{AB} 4	3	2	1
$\Sigma_g(2s)$–$\Sigma_u(2p)$	−0.135	−0.135	−0.149	−0.146
$(3p)$	0.334	0.380	0.423	0.361
$(4p)$	0.001	0.0002	0.008	0.039
$\Pi_u(2p)$	−0.334	−0.320	−0.283	−0.186
$(3p)$	0.905	0.826	0.711	0.522
$(4p)$	0.095	0.096	0.096	0.091
$\Sigma_g(3s)$–$\Sigma_u(2p)$	−0.026	−0.021	−0.016	−0.002
$(3p)$	−0.240	−0.249	−0.284	−0.275
$(4p)$	0.516	0.549	0.581	0.466
$\Pi_u(2p)$	−0.003	−0.002	−0.0004	−0.002
$(3p)$	−0.623	−0.595	−0.519	−0.344
$(4p)$	1.074	1.004	0.869	0.604
$\Sigma_u(2p)$–$\Pi_g(3d)$	0.552	0.435	0.344	0.352
$(4d)$	0.105	0.095	0.083	0.076
$(5d)$	0.039	0.038	0.034	0.030
$\Sigma_u(3p)$–$\Pi_g(3d)$	−0.0003	0.067	0.134	0.097
$(4d)$	0.424	0.320	0.232	0.261
$(5d)$	0.097	0.083	0.067	0.070
$\Sigma_u(4p)$–$\Pi_g(3d)$	−0.007	−0.017	−0.031	−0.027
$(4d)$	−0.026	0.086	0.208	0.162
$(5d)$	0.414	0.294	0.197	0.219
$\Pi_u(2p)$–$\Pi_g(3d)$	0.252	0.260	0.252	0.227
$(4d)$	0.005	0.012	0.024	0.033
$(5d)$	0.0007	0.003	0.007	0.012
$\Pi_u(3p)$–$\Pi_g(3d)$	−0.106	−0.100	−0.072	−0.029
$(4d)$	0.368	0.350	0.297	0.227
$(5d)$	0.016	0.028	0.040	0.043
$\Pi_u(4p)$–$\Pi_g(3d)$	−0.005	−0.002	−0.00001	−0.0016
$(4d)$	−0.195	−0.179	−0.131	−0.064
$(5d)$	0.458	0.414	0.327	0.239

approximation the radial $\tilde{\Pi}_u$ equation (and thus the energy) is identical to $\tilde{\Sigma}_u$. For these reasons, the approach of truncating the Hamiltonian is less precise than that of truncating the wavefunction. The situation is illustrated in Fig. 8 in which the one $\tilde{\Sigma}_u(=\tilde{\Pi}_u)$ energy is in worse accord with the exact energies than the previous zeroth-order results (cf. Fig. 3). The $\tilde{\Pi}_u$ are also *below* the exact results, which indicates that this approach is no longer variational with respect to the original Hamiltonian.

The identity of the radial $\tilde{\Sigma}_u$ and $\tilde{\Pi}_u$ equations ensures that each Σ_g–$\tilde{\Pi}_u$ oscillator will be double the corresponding Σ_g–$\tilde{\Sigma}_u$ value. Thus (3.2b) will automatically be satisfied if (3.2a) is. Table V contains the relevant results for the $\tilde{\Sigma}_u$ and $\tilde{\Pi}_u$ transitions. Comparing this with table III, one sees erratic improvement in the oscillator strengths notwithstanding the worsening of the energies at larger R_{AB}.

CONCLUSIONS

It has been shown in the case of $(H_2)^+$ that the relative partial wave expansion has many attractive features as a method of successive approximations. Most important is a strong convergence in going from one order of approximation to the next. It also has the properties of the orthogonality of the wavefunctions and the upper-boundedness of the energies; at the loss of the latter property one can modify the theory to guarantee the satisfaction of sum rules (equivalence of dipole length and dipole velocity forms of the transition integrals, etc.). In addition, the coordinates of the nonadiabatic expansion essentially eliminate the problem of evaluating multicenter integrals.

It is important to know which and to what extent these various features will be preserved in other possible applications. We are not in a position to answer all these questions at this time, but some points which are clear are appropriate to mention here. For instance, in applying this method to $(HeH)^{++}$ the convergence will not be as rapid as in $(H_2)^+$. That is so because, in the expansion of the wavefunction corresponding to (2.3), all relative angular momentum states must be included. (This in turn corresponds to the lack of symmetry about the origin and the loss of the quantum "numbers" gerade and ungerade in that system.) Thus, whereas in the $(H_2)^+$ system, the omitted term of the Σ_g wavefunction is depressed by a contrifugal barrier term $[-l(l+1)/r^2]$ with $l=2$, in the $(HeH)^{++}$ system the omitted term has $l=1$. This has the effect, for example, of eliminating Σ–Σ transitions

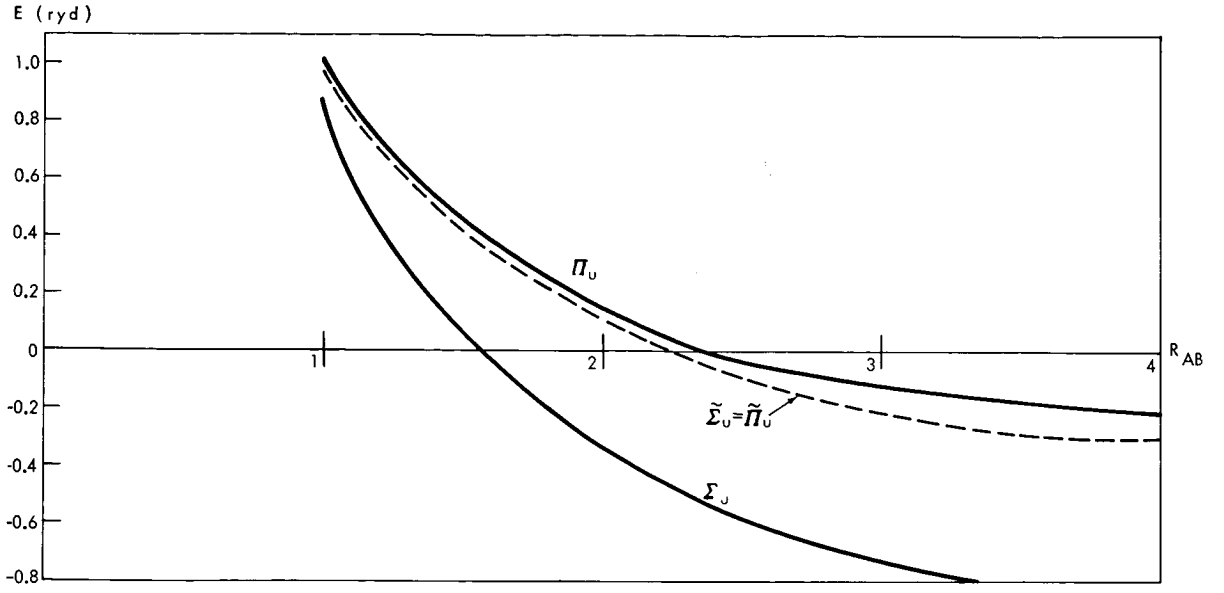


FIGURE 8.—Some modified zeroth energies.

in zeroth order whereas we know that exactly they are there.

In two electron diatomic molecules, which, as was stated in the Introduction, is our main objective, it is the goal of this approach to project a sequence of partial differential equations of significantly lower dimension (hopefully two dimensions) than the original Schrödinger equation such that solutions could be effected in a relatively precise and automatic way. We have in mind, specifically, a numerical method of integrating elliptic partial differential equations *noniteratively* which has recently been used in calculating the scattering of electrons from atomic hydrogen.[11]

Let it be noted that applications of the present idea to the ground state of H_2 have already been made.[12] Solutions in these cases were effected by expansion in the usual form of sums of separable

TABLE V.—*Modified Zeroth-Order Oscillator Strengths from the Ground State of* $(H_2)^+$

Transition	R_{AB}							
	4		3		2		1	
	Zeroth order	Exact	Zeroth order	Exact	Zeroth order	Exact	Zeroth order	Exact
$\Sigma_g(1s)$–$\tilde{\Sigma}_u$								
$2p$	0.325	0.238	0.314	0.289	0.289	0.319	0.226	0.269
$3p$	0.006	0.016	0.0108	0.004	0.019	0.0008	0.028	0.02
$4p$	0.0011	0.0046	0.0025	0.0015	0.0052	0.000055	0.009	0.006
Sum	0.3315	0.259	0.328	0.295	0.313	0.320	0.263	0.295
$\Sigma_g(1s)$–$\tilde{\Pi}_u(2p)$	0.650	0.46	0.628	0.47	0.479	0.46	0.454	0.39

functions. The restriction to lower order was incorporated by limiting the classes of functions which the basic set included. As such the fact that an equation simpler than the Schrödinger equation was being solved is hidden. When this equation is made manifest, then one is in more of a position to being the whole paraphernalia of dealing with partial differential equations to bear including numerical techniques. If accurate solutions can be obtained, then the properties of orthogonality upper-boundedness of the energies and/or satisfaction of the sum rules will again automatically be included.

ACKNOWLEDGMENTS

We should like to thank Theodore Stecher for stimulating us to look at these molecular problems by his many discussions of the associated astrophysical problems. The first-order problem for the Σ_g energies was programmed by R. Baxter and Patricia Egan.

REFERENCES

1. T. P. Stecher, Astrophys. J. **136,** 686 (1962). C. Zwaan, Bull. Astron. Inst. Neth. **16,** 225 (1962).
2. T. P. Stecher and J. E. Milligan, Astrophys J. **136,** 1 (1962); D. R. Bates, Monthly Notices Roy. Astron. Soc. **112,** 40 (1952).
3. P. Luke, R. Meyerott, and W. Clendenin, Phys. Rev. **85,** 401 (1952).
4. A. Temkin, Phys. Rev. **126,** 130 (1962).
5. D. R. Bates, K. Ledsham, and A. L. Stewart, Phil. Trans. Roy. Soc. (London) **A246,** 215 (1953).
6. D. R. Bates and T. R. Carson, Proc. Roy. Soc. (London) **A234,** 207 (1956).
7. M. Cohen and C. A. Coulson, Proc. Cambridge Phil. Soc. **57,** 96 (1961). We are grateful to Dr. A. Dalgarno for making this reference known to us. Some earlier applications of this idea are G. Huzinaga, Progr. Theoret. Phys. (Kyoto) **15,** 50 (1956); R. Gaspar, Acta. Phys. Hung. **7,** 151 (1957); T. Chen, J. Chem. Phys. **29,** 347 (1958); K. M. Howell and H. Shull, *ibid.* **30,** 627 (1959). The earlier treatments used various kinds of analytic approximation of the radial functions. Cohen and Coulson seem to be the first to have used a direct and accurate numerical solution of the radial equations. Previous treatments in general go under the name of single-center expansions.
8. M. Cohen, Proc. Cambridge Phil. Soc. **58,** 130 (1961).
9. For this reason any energy which goes negative will show a minimum in the nonadiabatic theory which may not actually be true. For such application this emphasizes the importance of joining onto an energy curve which gives the correct $R_{AB} \rightarrow \infty$ behavior.
10. We wish to thank T. P. Stecher for suggesting that we look at the sum rules. We are indebted to A. Dalgarno for pointing out to us that the $\Sigma_g(1s)$–$\Pi_u(2p)$ oscillator strengths for $R_{AB}=4$ actually exceeds the sum rule limit. This observation has directly led to the considerations of Sec. IV.
11. A. Temkin and E. Sullivan, Phys. Rev. **129,** 1250 (1963).
12. R. G. Parr and H. W. Joy, J. Chem. Phys. **26,** 424 (1957); S. Hagstrum and H. Shull, *ibid.* **30,** 1314 (1959). The latter paper contains references to many other papers along these lines.

NONSEPARABLE THEORY OF ELECTRON-HYDROGEN SCATTERING

A. TEMKIN AND D. E. HOOVER

Goddard Space Flight Center

INTRODUCTION*

The scattering of electrons from atomic hydrogen is a particularly appropriate subject for a book on machine methods of computation in physics. For—to our knowledge—it was on this venerable topic that the first application was made in physics of a large-scale computing machine. We refer to the work of Morse and Allis (1933) in which they used the Bush differential analyzer of M.I.T. to solve the integro-differential equations which arose in their generalization of the Hartree–Fock method to electron-atom scattering problems, as applied among others to hydrogen. This extension of the Hartree–Fock method was formally carried through by others as well as Morse and Allis. The unique aspect of the latter's work was to solve the resulting equations by "mechanical means" and obviate the necessity of making dubious analytic approximations. The important result of their calculations was found to be that the zero energy phase shifts approached π rather than zero radians. This constitutes the basis of the quantitative understanding of the phase shifts. In addition it provides a framework within which one can understand how the unincluded part of atomic polarization (distortion) can further reduce the electron-atom cross section. (This, in fact, characterizes much of the physical content of the present investigation.) The difficulty with the analytic approximation methods is that in one way or another they utilize a zero phase shift to compute a first-order correction which necessarily assumes that the final result is small. Thus they are completely incapable of dealing with the large phase shifts which usually occur in the case of low energy electron-atom scattering.

There may have been some querulousness at the time of the Morse and Allis paper because their results did not come out of some simple, elegant analytical formula. However, at the present time there can be no question of the validity of their work. It is, in fact, fair to say that the work of Morse and Allis inaugurates modern computational theoretical physics.

From a more general point of view, the equations that Morse and Allis solved were ultimately attempts to approximate solutions of the Schrödinger equation, a not unusual situation in atomic physics. Another of the important implications, then, of electronic computers for theoretical atomic physics is the fact that it allows one to think seriously of solving partial differential equations of two and perhaps more dimensions. (However, as we shall stress below, this is a task which must be approached with extreme caution.)

The Schrödinger equation for the S-wave scattering of electrons from hydrogen, can be written as (Morse and Feshbach, 1953)

$$\left\{-r_1^{-1}\frac{\partial^2}{\partial r_1^2}r_1-r_2^{-1}\frac{\partial^2}{\partial r_2^2}r_2-\left(\frac{1}{r_1^2}+\frac{1}{r_2^2}\right)\frac{1}{\sin\theta_{12}}\frac{\partial}{\partial\theta_{12}}\right.$$

$$\left.\times\sin\theta_{12}\frac{\partial}{\partial\theta_{12}}-\frac{2}{r_1}-\frac{2}{r_2}+\frac{2}{r_{12}}-E\right\}\Psi(r_1, r_2, \theta_{12})=0. \quad (1)$$

The S-wavefunction Ψ, which is the eigenfunction of total angular momentum zero, depends

* An outline of this method was given in Temkin (1960). A more complete discussion together with results appears in Temkin (1962).

only on three of the six coordinates, $\mathbf{r}_1$ and $\mathbf{r}_2$, of the problem. These can be chosen (as above) as the distances of the two electrons from the nucleus (assumed infinitely heavy and hence fixed), $\mathbf{r}_1$ and $\mathbf{r}_2$, and the angle between them, θ_{12}. The quantity r_{12} is

$$r_{12}=(r_1^2+r_2^2-2r_1r_2\cos\theta_{12})^{1/2}. \tag{2}$$

(Our units are lengths in Bohr radii, a_0, $a_0=\hbar^2/me^2=0.529\times10^{-8}$ cm and energy in rydbergs, 1 ry $=\hbar^2/2ma_0^2=13.6$ ev). The total energy consists of the energy of the hydrogen atom in its ground state (-1) plus the energy of the incident electron (k^2), thus

$$E=-1+k^2. \tag{3}$$

Let us remark at this point—in reference to the boundary conditions, which we shall immediately consider—that we shall restrict ourselves to energies below the threshhold for inelastic scattering,

$$E<-\tfrac{1}{4} \quad \text{or} \quad k^2<\tfrac{3}{4}.$$

In order to specify the boundary conditions succinctly, it is convenient to transform Eq. (1) slightly by introducing the function $\psi(r_1, r_2, \theta_{12})$ defined by

$$\psi(r_1, r_2, \theta_{12})\equiv r_1r_2\cdot\Psi(r_1, r_2, \theta_{12}).$$

Equation (1) then becomes

$$\left\{-\frac{\partial^2}{\partial r_1^2}-\frac{\partial^2}{\partial r_2^2}-\left(\frac{1}{r_1^2}+\frac{1}{r_2^2}\right)\frac{1}{\sin\theta_{12}}\frac{\partial}{\partial\theta_{12}}\sin\theta_{12}\frac{\partial}{\partial\theta_{12}}\right.$$

$$\left.-\frac{2}{r_1}-\frac{2}{r_2}+\frac{2}{r_{12}}-E\right\}\psi(r_1, r_2, \theta_{12})=0. \tag{4}$$

The properties of $\psi(r_1, r_2, \theta_{12})$ under the spatial exchange $\mathbf{r}_1\leftrightarrows\mathbf{r}_2$ are reflected by the condition

$$\psi(r_1, r_2, \theta_{12})=\pm\psi(r_2, r_1, \theta_{12}) \tag{5}$$

corresponding to singlet $(+)$ and triplet $(-)$ solutions. The existence of two solutions, symmetric and antisymmetric, is consistent with Eq. (4) since r_1 and r_2 enter symmetrically. In fact this means that the solutions with symmetry defined in Eq. (5) can be computed strictly in the region $r_1\geq r_2$ with an appropriate boundary condition. For the space symmetric solution (singlet) it is clear

$$\partial\psi/\partial n\,|_{r_1=r_2}=0, \tag{6a}$$

where $\partial/\partial n$ means the normal derivative (with θ_{12} constant) and for the space antisymmetric solution

$$\psi\,|_{r_1=r_2}=0. \tag{6b}$$

The advantage of ψ as opposed to Ψ is related to the fact that

$$\psi(r_1, 0, \theta_{12})=0. \tag{7}$$

In order to complete the mathematical statement of the problem, we need the boundary condition at infinity. This is given by

$$\lim_{r_1\to\infty}\psi(r_1, r_2, \theta_{12})=\sin(kr_1+\delta)R_{1s}(r_2). \tag{8}$$

$R_{1s}(r)$ is r times the ground-state wave function of hydrogen,

$$R_{1s}(r)=2re^{-r}.$$

The right hand side (rhs) of Eq. (8) then corresponds to an S-scattered electron at infinity, and the unperturbed hydrogen atom in its ground state centered at the origin. The restriction that $E<-\frac{1}{4}$ means that only the ground state R_{1s} fails to vanish in the asymptotic region defined by Eq. (8), and this is obviously a considerable simplification.

We have thus succeeded in expressing the S-wave scattering completely in terms of a partial differential equation with suitable boundary conditions. By introducing the function ψ we are able to give a well-defined boundary condition along $r_2=0$ in Eq. (7), and by considering the region $r_1\geq r_2$ only, we are able to convert the symmetry requirement, Eq. (5), into a boundary condition, Eq. (6). In summary, we are in principle trying to find *two* solutions of Eq. (4) both subject to the boundary conditions in Eq. (7) and Eq. (8) with singlet solutions obeying Eq. (6a) and the triplet obeying Eq. (6b).

The phase shifts δ are not at our disposal, but are determined by the remaining boundary conditions. This can be seen by supposing that one were actually contemplating a numerical solution of Eq. (4). In that case one would impose the boundary condition in Eq. (8) by letting

$$\psi(R, r_2, \theta_{12})=CR_{1s}(r_2)$$

for some suitably large $r_1=R$. This (Dirichlet) boundary condition combined with the other Dirichlet and/or Neumann boundary conditions around the other boundaries completely specifies

the solution of Eq. (4), which is an elliptic equation. For different values of C, we would obtain solutions that differ from each other only by the values of the constant C, corresponding to the homogeneity of Eq. (4). A value of δ deduced from any solution will obviously be independent of C. The uniqueness of δ then follows from the (assumed) uniqueness of the solution for a given C.

In mentioning a numerical solution, we want to emphasize that this is not what we have done in solving the S-wave scattering problem. It may be that computer capabilities are such that the possibility of a numerical solution exists; however, we doubt it. Before describing what we have done, we should like to interject a word of philosophy concerning the question of numerical solution. And that is, one must exercise considerable caution before embarking on a vast program of direct numerical solution of a complicated problem, even if apparently well-defined. This is particularly true of partial differential equations. Of course, in any problem, a time does arrive when numerical work is necessary. In our opinion, when the work we shall describe was done, time was not ripe for the direct numerical integration of this equation. Specific reasons will be given subsequently. It is our hope that as a result of this work, numerical integration will become a more sensible project.

REDUCTION OF THE S-WAVE EQUATION

Having renounced attempts at direct numerical solution of Eq. (4), we must proceed to simplify it to manageable proportions. The basic step in this program is the following expansion of $\psi(r_1, r_2, \theta_{12})$ †

$$\psi(r_1, r_2, \theta_{12}) = \sum_{l=0}^{\infty} (2l+1)^{1/2} \Phi_l(r_1, r_2) P_l(\cos\theta_{12}). \tag{9}$$

$P_l(\cos\theta_{12})$ are the Legendre polynomials of $\cos\theta_{12}$, and $\Phi_l(r_1, r_2)$ are functions determined by substituting Eq. (9) into Eq. (4). This substitution yields an infinite set of coupled two-dimensional partial differential equations,

$$\left\{\frac{\partial^2}{\partial r_1^2}+\frac{\partial^2}{\partial r_2^2} - l(l+1)\left(\frac{1}{r_1^2}+\frac{1}{r_2^2}\right)+E+\frac{2}{r_1}+\frac{2}{r_2} - M_{ll}\right\}\Phi_l(r_1, r_2) = \sum_{m=0}^{\infty}{}' M_{lm}\Phi_m(r_1, r_2), \tag{10}$$

where the M_{lm} are also functions of r_1 and r_2,

$$M_{lm} = [(2l+1)(2m+1)]^{1/2} \sum_{n=0}^{l+m} \frac{r_2^n}{r_1^{n+1}} \int_0^{\pi} P_l(\cos\theta) P_m(\cos\theta) P_n(\cos\theta) \sin\theta \, d\theta. \tag{11}$$

The boundary conditions, Eqs. (6) to (8), in terms of the functions Φ_l become

$$(\partial/\partial n)\Phi_l(r_1, r_2) \,|_{r_1=r_2} = 0, \qquad \text{singlet} \tag{12a}$$

$$\Phi_l(r_1, r_2) \,|_{r_1=r_2} = 0, \qquad \text{triplet} \tag{12b}$$

$$\Phi_l(r_1, 0) = 0 \tag{13}$$

$$\lim_{r_1\to\infty} \Phi_l(r_1, r_2) = \sin(kr_1+\delta) R_{1s}(r_2) \delta_{l0}, \tag{14}$$

where δ_{l0} is the Kronecker delta. The last equation says that all the Φ_l, but Φ_0, approach zero asymptotically. This condition is implicit in Eq. (8), the rhs of which is independent of θ_{12}. *However*, for $l>0$ *the* Φ_l *approach zero slowly*, as we shall now show. Let us consider the $l>0$ of Eqs. (10) in the adiabatic

† This expansion as well as the technic of solution of the zeroth-order problem was first used in a bound state problem by Luke et al. (1952). The idea was suggested by G. Breit. The original application was to calculate the energy of an excited S-state of singly ionized lithium. The present method can be described as an extension of that approach to the scattering problem. For the reason that the scattering wave function has a basically different and more important adiabatic behavior which this method very naturally isolates and allows to be analytically determined, we feel that the method applies better to the scattering problem.

region ($r_1 \gg r_2$, $r_1 \gg 1$). The equations reduce to

$$\left(\frac{\partial^2}{\partial {r_1}^2}+\frac{\partial^2}{\partial {r_2}^2}-\frac{l(l+1)}{{r_2}^2}+E+\frac{2}{r_2}\right)\Phi_l^{(\mathrm{adiab})}=\frac{2}{(2l+1)^{1/2}}\frac{{r_2}^l}{{r_1}^{l+1}}\sin(kr_1+\delta)R_{1s}(r_2).$$

The solution of this equation, neglecting terms $r_1^{-\mu}$, $\mu > l+1$, is

$$\Phi_l^{(\mathrm{adiab})}=\frac{-2}{(2l+1)^{1/2}}\frac{\sin(kr_1+\delta)}{{r_1}^{l+1}}e^{-r_2}\left(\frac{{r_2}^{l+2}}{l+1}+\frac{{r_2}^{l+1}}{l}\right). \tag{15}$$

From the foregoing we see that $\Phi_l \to 0$ as ${r_1}^{-l-1}$ which for small l is very much slower than the exponential decay that occurs in bound state wave functions.

The actual form of $\Phi_l^{(\mathrm{adiab})}$ will be necessary in Section IV, wherein we consider the higher order corrections; but in connection with the problem of numerical integration, we remark that the slow vanishing of Φ_l provides a concrete example of a major difficulty in obtaining an adequate direct numerical solution of the three-dimensional equation for these boundary conditions. Because of problems of storage and time, one is forced to stop a numerical integration at a point $r_1 = R$ which cannot be too large since the number of mesh points goes up as the cube of the linear dimension. However, unless R is very large, the asymptotic form in Eq. (8) will be significantly altered by the slowly vanishing angular dependent terms implicit in Eq. (11). In order practically to include them in a numerical calculation, one would probably have to do so analytically.

This consideration also has a practical relevance to methods that are actually in use. Variational techniques when applied to scattering problems must explicitly include the nonvanishing term contained in Eq. (8). However, beyond that, it is customary to use terms that die out exponentially. The present considerations show that such functions cannot completely take account of the long-range induced polarization, which is manifested in the functions $\Phi_l^{(\mathrm{adiab})}$. Actual calculations show that the adiabatic behavior gives direct contributions that are quantitatively important for low energies. (Temkin, 1961a).

To return to the formal development, let us write explicitly the $l=0$ equation of the set of Eqs. (10);

$$\left(\frac{\partial^2}{\partial {r_1}^2}+\frac{\partial^2}{\partial {r_2}^2}+E+\frac{2}{r_2}\right)\Phi_0(r_1, r_2)=\sum_{m=1}^{\infty}\frac{2}{(2m+1)^{1/2}}\frac{{r_2}^m}{{r_1}^{m+1}}\Phi_m. \tag{16}$$

We define a zeroth-order problem by neglecting the rhs of Eq. (16)

$$\left(\frac{\partial^2}{\partial {r_1}^2}+\frac{\partial^2}{\partial {r_2}^2}+E+\frac{2}{e_2}\right)\Phi_0^{(0)}(r_1, r_2)=0. \tag{17}$$

The solution of this equation, $\Phi_0^{(0)}$, of course, is different from Φ_0. Nevertheless, we should like it to be as much like Φ_0 as possible. Such a solution can be obtained by imposing the boundary conditions

$$\Phi_0^{(0)}(r_1, 0)=0 \tag{18}$$

$$(\partial/\partial n)\Phi_0^{(0)}(r_1, r_2)\,|_{r_1=r_2}=0, \qquad \text{singlet} \tag{19a}$$

$$\Phi_0^{(0)}(r_1, r_2)\,|_{r_1=r_2}=0, \qquad \text{triplet} \tag{19b}$$

$$\lim_{r_1\to\infty}\Phi_0^{(0)}(r_1, r_2)=\sin(kr_1+\delta_0)R_{1s}(r_2). \tag{20}$$

By virtue of the other boundary conditions, δ_0, like δ is not arbitrary. In fact, the determination of δ_0 and $\Phi_0^{(0)}$, which together constitutes the zeroth-order problem, is the major numerical problem with which the bulk of this report will be concerned. A significant feature of the zeroth-order problem stems from the fact that one can derive an exact relation between δ and δ_0. In essence this relation contains the epitome of the physics and mathematics involved in this S-wave scattering problem. We shall conclude this formal section with the derivation of that relation.

Multiply Eq. (17) on the left by Φ_0 and Eq. (16) by $\Phi_0^{(0)}$; subtract and integrate over all r_1 and r_2. This gives

$$\int_0^\infty dr_1 \int_0^{r_1} dr_2 (\Phi_0 \Delta_{12} \Phi_0^{(0)} - \Phi_0^{(0)} \Delta_{12} \Phi_0) = -\int_0^\infty \int_0^{r_1} \Phi_0^{(0)} \sum_{m=1}^{\infty} \frac{2}{(2m+1)^{1/2}} \Phi_m \, dr_1 \, dr_2. \tag{21}$$

The operator Δ_{12} is the two-dimensional Laplacian

$$\Delta_{12} \equiv (\partial^2/\partial r_1^2) + (\partial^2/\partial r_2^2). \tag{22}$$

Thus one can convert the lhs of Eq. (21) to a line integral using Green's theorem,

$$\int_0^\infty \int_0^{r_1} (\Phi_0 \Delta_{12} \Phi_0^{(0)} - \Phi_0^{(0)} \Delta_{12} \Phi_0) \, dr_1 \, dr_2 = \int_s (\Phi_0 (\partial/\partial n) \Phi_0^{(0)} - \Phi_0^{(0)} (\partial/\partial n) \Phi_0) \, ds.$$

The boundary s goes around the $r_1 \geq r_2$ triangle as shown in Fig. 1. Along s_1, the integral is zero by virtue of Eqs. (13) and (18); along s_3 it is zero, in the singlet by Eqs. (12a) and (19a), in the triplet case by

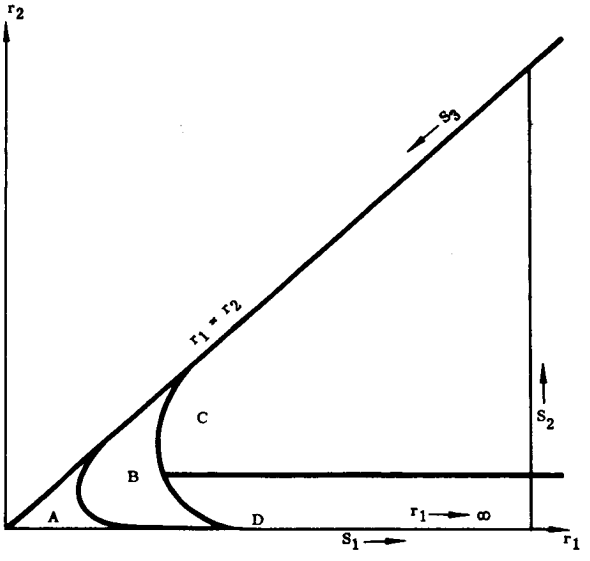


FIGURE 1.—Significant regions and boundaries of the $r_1 > r_2$ triangle.

Eqs. (12b) and (19b). We are left with

$$\int_s [\Phi_0 (\partial/\partial n) \Phi_0^{(0)} - \Phi_0^{(0)} (\partial/\partial n) \Phi_0] \, ds = \int_{s_2} [\Phi_0 (\partial/\partial n) \Phi_0^{(0)} - \Phi_0^{(0)} (\partial/\partial n) \Phi_0] \, ds_2.$$

Here we use the fact that $s_2 = r_2$, $\partial/\partial n = \partial/\partial r_1$, and $r_1 \to \infty$ so that we can use Eqs. (14) and (20) for Φ_0 and $\Phi_0^{(0)}$, respectively. We thus get

$$\begin{aligned} \int [\Phi_0 (\partial/\partial n) \Phi_0^{(0)} - \Phi_0^{(0)} (\partial/\partial n) \Phi_0] \, ds_2 &= \lim_{r_1 \to \infty} \int_0^\infty [\Phi_0 (\partial/\partial r_1) \Phi_0^{(0)} - \Phi_0^{(0)} (\partial/\partial r_1) \Phi_0] \, dr_2 \\ &= k[\sin(kr_1 + \delta) \cos(kr_1 + \delta_0) \\ &\qquad - \cos(kr_1 + \delta) \sin(kr_1 + \delta_0)] \int_0^\infty R_{1s}^2(r_2) \, dr_2 \\ &= k \sin(\delta - \delta_0), \end{aligned}$$

where we have chosen $R_{1s}(r)$ to be normalized. Combining the foregoing with the rhs of Eq. (21) we obtain the basic equation

$$\sin(\delta-\delta_0) = -k^{-1}\sum_{m=1}^{\infty}\frac{2}{(2m+1)^{1/2}}\int_0^{\infty}dr_1\int_0^{r_1}dr_2\Phi_0^{(0)}\frac{r_2^m}{r_1^{m+1}}\Phi_m. \tag{23}$$

The corrections to the zeroth order δ_0 are given by the terms on the rhs of Eq. (23), whose chief property, mathematically speaking, is that they are rapidly convergent. The arguments for this convergence are for the present heuristic, but nevertheless cogent. The triangular configuration space of each double integral on the rhs of Eq. (23) can be divided into four regions as shown schematically in Fig. 1. Region D is the adiabatic region where the function Φ_l becomes identical with $\Phi_l^{(\mathrm{adiab})}$ of Eq. (15). The convergence of that part of the integral as a function of l can be explicitly demonstrated by using the known forms of $\Phi_l^{(\mathrm{adiab})}$ and $\Phi_0^{(0)}$, the latter being given there by Eq. (20). In region C all functions vanish exponentially as $\exp[+(|E|/2)^{1/2}(r_1+r_2)]$, so that this contribution is negligible in all cases. In region A, we see on referring to Eq. (10) that the effective centrifugal barrier

$$-l(l+1)(1/r_1^2+1/r_2^2), \quad \text{for} \quad l>0$$

pushes the corresponding Φ_l out of that region. This pushing out increases as a function of l, so that region A gradually swallows up region B, the latter however, giving an important quantitative contribution for small l.

Equation (23) contains, implicitly, physical statements concerning the electron-hydrogen scattering problem. In particular, δ_0 is the phase shift of a very rudimentary three-body problem, namely, one in which the incoming electron is completely shielded from the nucleus as long as it is outside of the orbital electron, but sees the complete nuclear charge when it is inside. Although this shielding approximation may appear crude, it should be noted that correlations between the electrons are still retained. Mathematically, this situation is reflected in the fact that although the zeroth-order equation is separable, the boundary conditions (along $r_1=r_2$) are not. It will turn out that the correlations of the zeroth-order problem are the essential ones of the problem as a whole. The terms on the rhs of Eq. (23) correspond to induced multipole distortions of the atom which are obviously not taken into account in δ_0. Here too, however, in the intermediate region of r_1 and r_2 (region B) there are nonadiabatic effects which mathematically are reflected by the nonseparable nature of the functions Φ_m as well as $\Phi_0^{(0)}$ in that region. These nonseparable effects can fully be taken into account in this treatment by virtue of the two dimensionality of the equations. The nonadiabatic behavior of $\Phi_l(l>0)$ is not as important quantitatively as those associated with $\Phi_0^{(0)}$, since a large part of the integrals comes from the adiabatic region (region D), at least for low energies, where the functions Φ_l are mainly separable. This means that function $\Phi_0^{(0)}$ should be known more precisely than the functions Φ_l. In the next section we shall detail our rather elaborate and somewhat novel attempt (in the context of a scattering problem) to obtain $\Phi_0^{(0)}$. In Section IV we shall describe what we have done for the higher Φ_l.

SOLUTION OF THE ZEROTH-ORDER PROBLEM

The zeroth-order problem has been developed and presented in the preceding Section. This section deals with our attempt to solve that problem. See footnote, p. 1603.

Recapitulating, we seek two solutions of

$$\left(\frac{\partial^2}{\partial r_1^2}+\frac{\partial^2}{\partial r_2^2}+E+\frac{2}{r_2}\right)\Phi_0^{(0)}(r_1, r_2)=0 \tag{17}$$

under the constraints

$$\Phi_0^{(0)}(r_1, 0)=0 \tag{18}$$

$$\lim_{r_1\to\infty}\Phi_0^{(0)}(r_1, r_2)=\sin(kr_1+\delta_0)R_{1s}(r_2), \tag{20}$$

and

$$(\partial/\partial n)\Phi_0^{(0)}(r_1, r_2)\,|_{r_1=r_2}=0, \quad \text{singlet} \tag{19a}$$

$$\Phi_0^{(0)}(r_1, r_2)\,|_{r_1=r_2}=0, \quad \text{triplet} \tag{19b}$$

where, as has been pointed out earlier, we are investigating the region

$$r_2 \leq r_1.$$

Seeing that Eq. (17) is separable, we set about to find solutions which are the sum of all possible solutions having the form of a product $f(r_1)\cdot g(r_2)$. By substituting into Eq. (17) and on using the usual technique of separation of variables, one gets two ordinary differential equations. The first is

$$[g(r_2)]^{-1}[\partial^2 g(r_2)/\partial r_2^2]+(2/r_2)=\text{constant}. \quad (24)$$

It will become clear that $g(r_2)$ must be bounded at infinity and that $g(0)=0$. The second equation is

$$[f(r_1)]^{-1}[\partial^2 f(r_1)/\partial r_1^2]+E+\text{constant}=0. \quad (25)$$

The function $f(r_1)$ needs only to be bounded at infinity and regular at $r_1=0$.

The equations

$$\left(\frac{d^2}{dr_2^2}+\frac{2}{r_2}-\frac{3}{n^2}\right)g(r_2)=0$$

are the well-known equations for the discrete s-states of the hydrogen atom. We begin therefore by making $1/n^2$ one choice for our constant. We let $g(r_2)=R_{ns}(r_2)$ which is the usual nomenclature for r_2 times the radial hydrogenic wave function.

We can also choose a negative constant, $-p^2$, and get

$$[(d^2/dr_2^2)+(2/r_2)+p^2]u_p(r_2)=0. \quad (25a)$$

These are the equations for the continuum s-states of hydrogen corresponding to the hyperbolic orbits of an electron in the field of a unit charge, the solutions to which are also known.

These functions $R_{ns}(r_2)$ and $u_p(r_2)$ are zero at the origin and bounded at infinity as required by Eq. (24).

We next consider the equation

$$[f(r_1)]^{-1}[\partial^2 f(r_1)/\partial r_1^2]+E+(1/n^2)=0,$$

and referring to Section I, we write

$$\{(\partial^2/\partial r_1^2)-[1-k^2-(1/n^2)]\}f(r_1)=0, \qquad k^2<\tfrac{3}{4}.$$

By way of emphasis, $k^2<\frac{3}{4}$ implies, when $n=1$, that $(1-k^2-1/n^2)<0$. But when $n\geq 2$, $(1-k^2-1/n^2)$ is positive. We treat these two cases separately.

If $n=1$,

$$[(d^2/dr_1^2)+k^2]f(r_1)=0$$

$$\therefore f(r_1)=A\sin(kr_1+\delta_0).$$

If $n>1$, letting $(1-k^2-1/n^2)=\kappa_n^2$, we get

$$[(d^2/dr_1^2)-\kappa_n^2]f(r_1)=0$$

$$\therefore f(r_1)=C_{n1}\exp(-\kappa_{n1}r_1)+C_{n2}\exp(\kappa_{n2}r_1).$$

In order that $f(r_1)$ be bounded at infinity, C_{n2} must be set equal to zero.

For our constant, $-p^2$,

$$[f(r_1)]^{-1}[\partial^2 f(r_1)/\partial r_1^2]+E-p^2=0,$$

so that

$$[(d^2/dr_1^2)-(1-k^2+p^2)]f(r_1)=0,$$

where we define $(1-k^2+p^2)=\kappa_p^2$. Then

$$[(d^2/dr_1^2)-\kappa_p^2]f(r_1)=0$$

and

$$\therefore f(r_1)=C_{p1}\exp(-\kappa_{p1}r_1)+C_{p2}\exp(\kappa_{p2}r_1).$$

The constant C_{p2} must also be set equal to zero.

We can now write down our answer as

$$\Phi_0^{(0)}=\sin(kr_1+\delta_0)R_{1s}(r_2)+\sum_i C_i\exp(-\kappa_i r_1)R_{is}(r_2)+\int_p C_p\exp(-\kappa_p r_1)u_p(r_2)\,dp, \quad (26)$$

where the constant A has been set equal to 1 by Eq. (20). The integral sign indicates a sum of solutions where solutions are dense. For computational purposes, we make only finite samplings of the continuum states, so that hereafter in this section the integral sign will be replaced by a summation sign. The constraining relations imposed on $R_{ns}(r_2)$ and $u_p(r_2)$ plus the restriction $r_2\leq r_1$ guarantees that the above expansion of $\Phi_0^{(0)}$ automatically fulfills the boundary conditions in Eqs. (18) and (20).

We satisfy the remaining constraints by stipulating that the constants δ_0 and C_m assume the

particular values such that

$$\Phi_0^{(0)}(r_1, r_2)\ |_{r_1=r_2}=0 \qquad (19b)$$

in the triplet case, and

$$(\partial/\partial n)\Phi_0^{(0)}(r_1, r_2)\ |_{r_1=r_2}=0 \qquad (19a)$$

in the singlet case are satisfied. As long as we restrict ourselves to a discrete subset of the continuum of $u_p(r_2)$ states, Eqs. (19) can only be satisfied approximately.

Let

$$I_T\equiv\int_0^\infty[\Phi_0^{(0)}(r_1, r_2)\ |_{r_1=r_2}]^2\ dr_1 \qquad (27)$$

and

$$I_S\equiv\int_0^\infty[(\partial/\partial n)\Phi_0^{(0)}(r_1, r_2)\ |_{r_1=r_2}]^2\ dr_1. \qquad (28)$$

The technique of solution, then, is to determine the C_i's and δ_0 by minimizing the I_T and I_S, when Eq. (26) is formally inserted into them, for the triplet and singlet results, respectively. We shall treat the derivation of the resulting triplet formulas in some detail.

The substitution of Eq. (26) into Eq. (27) gives

$$I_T\equiv\int_0^\infty[\sin(kr_1+\delta_0)R_{1s}(r_2)+\sum_i C_i\exp(-\kappa_i r_1)\times R_{is}(r_2)+\sum_{p_i}C_{p_i}\exp(-\kappa_{p_i}r_1)u_{p_i}(r_2)]^2_{r_1=r_2}\ dr_1.$$

The sum of terms made up of the product of the first term by any element of the terms under summation signs we group together as $2\sum_j C_jY_j$. The sum of the squares of the elements of the two sum terms, we call $\sum_j C_j^2(\mathrm{ME})_{jj}$. The cross products with factors among the two sums we group together and call $2\sum_{j>i}\sum C_iC_j(\mathrm{ME})_{ij}$. Using this symbolism we can write

$$I_T=\int_0^\infty\sin^2(kr+\delta_0)R_{1s}^2(r)\ dr+2\sum_j C_jY_j+\sum_j C_j^2(\mathrm{ME})_{jj}+2\sum_{j>i}\sum C_iC_j(\mathrm{ME})_{ij},$$

where I_T is a function of our C_i's and δ_0 and

$$(\mathrm{ME})_{ij}=\int_0^\infty\exp[-(\kappa_i+\kappa_j)r]R_{is}(r)R_{js}(r)\ dr.$$

This matrix element is written as if both states were discrete states. Actually, this is one of three types of matrix elements which can occur, the other two being discrete-continuum and continuum-continuum types. The latter can be written

$$(\mathrm{ME})_{Ip_1,Ip_2}=\int_0^\infty\exp[-(\kappa_{p_1}+\kappa_{p_2})r]\times u_{p_1}(r)u_{p_2}(r)\ dr. \qquad (29)$$

This is perhaps the most involved of various integrals containing continuum Coulomb wave functions which are needed. The integration of Eq. (29) is carried out in Appendix I. A compendium of the nontrivial integrals is given in Temkin (1962).

Numerically, we look for our minimizing constants in the simultaneous solutions of the equations

$$-\partial I_T/\partial C_i=0, \qquad i=1, \cdots, N_3 \qquad (30)$$

$$\partial I_T/\partial\delta_0=0.$$

The first equations of this system are

$$2Y_i+2C_i(\mathrm{ME})_{ii}+2\sum_{j\neq i}C_j(\mathrm{ME})_{ij}=0, \qquad i=1, \cdots, N_3$$

or more simply

$$Y_i+\sum_j C_j(\mathrm{ME})_{ij}=0. \qquad i=1, \cdots, N_3$$

This can be considered a system of N_3 equations for the N_3 unknown C_i's which then lends itself obviously to a solution by Cramer's rule. However, the Y_i are actually functions of δ_0. For the discrete case (and analogously for the continuum cases)

$$Y_i=\int_0^\infty\exp(-\kappa_i r)R_{is}(r)R_{1s}(r)\sin(kr+\delta_0)\ dr.$$

Rewritten this becomes

$$Y_i=\cos\delta_0\int_0^\infty \exp(-\kappa_i r)\,R_{is}(r)\,R_{1s}(r)\,\sin(kr)\,dr$$

$$+\sin\delta_0\int_0^\infty \exp(-\kappa_i r)\,R_{is}(r)\,R_{1s}(r)\,\cos(kr)\,dr, \tag{31}$$

which we write symbolically as

$$Y_i=\mathfrak{N}_{si}\delta_0+\mathfrak{N}_{ci}\sin\delta_0. \tag{32}$$

Here $\mathfrak{N}_{si}$ and $\mathfrak{N}_{ci}$ are defined as inferred in Eq. (31).

$$\therefore C_i=(-1/\mathrm{DET})[\mathfrak{D}_s^{(i)}\cos\delta_0+\mathfrak{D}_c^{(i)}\sin\delta_0], \tag{33}$$

with DET being the determinant of the matrix elements.

$$\mathrm{DET}=\begin{vmatrix} \mathrm{ME}_{11} & \mathrm{ME}_{12} & \cdots & \mathrm{ME}_{1N3} \\ \mathrm{ME}_{21} & \mathrm{ME}_{22} & \cdots & \mathrm{ME}_{2N3} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \mathrm{ME}_{N31} & \mathrm{ME}_{N32} & \cdots & \mathrm{ME}_{N3N3} \end{vmatrix}$$

and $\mathfrak{D}_s^{(i)}$ and $\mathfrak{D}_c^{(i)}$ are obtained from DET by replacing the ith row by the row vectors $\mathfrak{N}_{s1}, \cdots, \mathfrak{N}_{sN3}$ and $\mathfrak{N}_{c1}, \cdots, \mathfrak{N}_{cN3}$, respectively.

The last equation of our simultaneous system Eq. (30) is

$$2\int_0^\infty \sin(kr+\delta_0)\,\cos(kr+\delta_0)\,R_{1s}^2(r)\,dr+2\sum_i C_i[-\mathfrak{N}_{si}\sin\delta_0+\mathfrak{N}_{ci}\cos\delta_0]=0,$$

wherein we have made use of Eq. (32). The familiar trigonometric identities and Eq. (33) allow this equation to be written

$$\mathfrak{N}_{2s}\cos2\delta_0+\mathfrak{N}_{2c}\sin2\delta_0+2\cos\delta_0\sum_i\{(-1/\mathrm{DET})[\mathfrak{D}_s^{(i)}\cos\delta_0+\mathfrak{D}_c^{(i)}(\sin\delta_0]\mathfrak{N}_{ci}\}$$

$$-2\sin\delta_0\sum_i\{(-1/\mathrm{DET})[\mathfrak{D}_s^{(i)}\cos\delta_0+\mathfrak{D}_c^{(i)}\sin\delta_0]\mathfrak{N}_{si}\}=0,$$

with

$$\mathfrak{N}_{2s}=\int_0^\infty \sin(2kr)\,R_{1s}^2(r)\,dr$$

and

$$\mathfrak{N}_{2c}=\int_0^\infty \cos(2kr)\,R_{1s}^2(r)\,dr.$$

A regrouping yields

$$\mathfrak{N}_{2s}\cos2\delta_0+\mathfrak{N}_{2c}\sin2\delta_0-(2/\mathrm{DET})\sin\delta_0\cos\delta_0\sum_i(\mathfrak{D}_c^{(i)}\mathfrak{N}_{ci}-\mathfrak{D}_s^{(i)}\mathfrak{N}_{si})$$

$$\times(-2/\mathrm{DET})\sum_i(\cos^2\delta_0\mathfrak{D}_s^{(i)}\mathfrak{N}_{ci}-\sin^2\delta_0\mathfrak{D}_c^{(i)}\mathfrak{N}_{si})=0.$$

The known symmetry of the determinant of our matrix elements now leads to the useful equality

$$\sum_i\mathfrak{D}_s^{(i)}\mathfrak{N}_{ci}=\sum_i\mathfrak{D}_c^{(i)}\mathfrak{N}_{si}. \tag{33a}$$

Thus

$$\mathfrak{N}_{2s}\cos2\delta_0+\mathfrak{N}_{2c}\sin2\delta_0-(\sin2\delta_0/\mathrm{DET})\sum_i(\mathfrak{D}_c^{(i)}\mathfrak{N}_{ci}-\mathfrak{D}_s^{(i)}\mathfrak{N}_{si})\,(-2/\mathrm{DET})\cos2\delta_0\sum_i\mathfrak{D}_s^{(i)}\mathfrak{N}_{ci}=0.$$

The solving of this last equation for $\tan 2\delta_0$ yields the final result

$$\tan 2\delta_0 = \frac{-\mathfrak{N}_{2s} + (2/\mathrm{DET}) \sum_i \mathfrak{D}_s^{(i)} \mathfrak{N}_{ci}}{\mathfrak{N}_{2c} - (1/\mathrm{DET}) \sum_i (\mathfrak{D}_c^{(i)} \mathfrak{N}_{ci} - \mathfrak{D}_s^{(i)} \mathfrak{N}_{si})}$$

for the triplet case.

For the singlet case, we define the total square value of the normal derivative of our function $\Phi_0^{(0)}$ as

$$I_S = \int_0^\infty [(\partial/\partial n)\, \Phi_0^{(0)} \,|_{r_1=r_2}]^2 \, dr_1. \tag{28}$$

In this context, the operator

$$\partial/\partial n \equiv -(\partial/\partial r_1) + (\partial/\partial r_2)$$

which is $\sqrt{2}$ times the normal derivative. The substitution of the finite expansion for $\Phi_0^{(0)}$ in Eq. (28) yields

$$I_S = \int_0^\infty [-k \cos(kr+\delta_0) + \sin(kr+\delta_0)\,(r^{-1}-1)]^2 R_{1s}^2(r)\, dr + 2\sum_j C_j[-\mu_j \cos\delta_0 - \nu_j \sin\delta_0] + \sum_j C_j^2 (\mathrm{SME})_{jj} + 2\sum_{j>i}\sum C_i C_j (\mathrm{SME})_{ij},$$

where

$$\nu_j = (SN)_{crj} - (\mathrm{SN})_{cj} - k(\mathrm{SN})_{srj}$$

$$\mu_j = (SN)_{srj} - (SN)_{sj} + k(SN)_{crj}$$

and

$$(SN)_{crj} = \int_0^\infty \cos(kr_1)\, R_{1s}(r_1)\, (\partial/\partial n)[\exp(-\kappa r_1) g(r_2)]\,|_{r_1=r_2}\, dr_1$$

$$(SN)_{srj} = \int_0^\infty \sin(kr_1)\, R_{1s}(r_1)\, (\partial/\partial n)[\exp(-\kappa r_1) g(r_2)]\,|_{r_1=r_2}\, dr_1$$

$$(SN)_{cj} = \int_0^\infty r_1^{-1} \cos(kr_1)\, R_{1s}(r_1)\, (\partial/\partial n)[\exp(-\kappa r_1) g(r_2)]\,|_{r_1=r_2}\, dr_1$$

$$(SN)_{sj} = \int_0^\infty r_1^{-1} \sin(kr_1)\, R_{1s}(r_1)\, (\partial/\partial n)[\exp(-\kappa r_1) g(r_2)]\,|_{r_1=r_2}\, dr_1$$

The κ and $g(r_2)$ are either κ_j and $R_{js}(r_2)$ or κ_{pj} and $u_{pj}(r_2)$, respectively, depending on whether the subscript j of the left members of the last four equations refers to a discrete or continuum state. Similarly,

$$(\mathrm{SME})_{ij} = \int_0^\infty (\partial/\partial n)[\exp(-\kappa r_1) g(r_2)]_i (\partial/\partial n)[\exp(-\kappa r_1) g(r_2)]_j \,|_{r_1=r_2}\, dr_1,$$

which likewise can be a discrete-discrete state, discrete-continuum, or continuum-continuum state.

To determine the C_i's and δ_0 that minimize I_S, we again differentiate with respect to these quantities and solve the resulting system of equations simultaneously for them.

This yields

$$C_i = [D_\mu^{(i)} \cos\delta_0 + D_\nu^{(i)} \sin\delta_0 / \mathrm{DETS}]$$

and

$$\tan 2\delta_0 = \frac{\frac{1}{2}A_s - kB_c - 2/\mathrm{DETS}\sum_j \nu_j \mathfrak{D}_\mu^{(j)}}{-\frac{1}{2}A_c - kB_s - 1/\mathrm{DETS}\sum_j (\mu_j \mathfrak{D}_\mu^{(j)} - \nu_j \mathfrak{D}_\nu^{(j)})}.$$

The determinants involved have definitions analogous to those in the triplet results.

Also

$$A_s = 8\int_0^\infty \sin(2kr)[1-2r+(1-k^2)r^2]e^{-2r}\,dr,$$

$$A_c = 8\int_0^\infty \cos(2kr)[1-2r+(1-k^2)r^2]e^{-2r}\,dr,$$

$$B_s = 8\int_0^\infty \sin(2kr)r(1-r)e^{-2r}\,dr,$$

$$B_c = 8\int_0^\infty \cos(2kr)r(1-r)e^{-2r}\,dr.$$

The equality

$$\sum_j \nu_j \mathfrak{D}_\mu^{(j)} = \sum_j \mu_j \mathfrak{D}_\nu^{(j)}$$

plays an analogous role to Eq. (33a) in the derivation.

Because of the periodicity of the tangent, the above formulas determine $2\delta_0$ to within π. Examination of the $2\delta_0$'s (modulo 2π) determines which value of $2\delta_0$ minimizes I_T. This still leaves δ_0 undetermined modulo π.

It has been described by one of us (Temkin, 1961b) how δ_0 can be obtained modulo 2π. In brief, the method consists of defining a one-particle scattered orbital which, in this case, would be

$$u(r_1) = \int_0^\infty R_{1s}(r_2)\Phi_0^{(0)}(r_1, r_2)\,dr_2. \tag{34}$$

Phase shifts are to be found from $u(r_1)$ as in any one-body problem. In particular one requires that

$$u'(0) > 0. \tag{35}$$

This, then, is the condition from which δ_0 can be found to modulo 2π. In fact, this can be used to define an absolute value of the phase shift. This means that one has to evaluate $u'(0)$ from Eq. (34) using the two values of δ_0 (modulo 2π) and see which one satisfied the condition of Eq. (35).

Actual calculations show in both singlet and triplet cases that $\delta_0 \to \pi$ from below as $k \to 0$. In performing the integral in Eq. (34) one needs values of $\Phi_0^{(0)}$ for $r_2 > r_1$. This point has also been discussed in the previously mentioned article wherein it is shown for the triplet case that

$$\Phi_0^{(0)}(r_1, r_2)\,|_{r_1 \leq r_2} = -\Phi_0^{(0)}(r_2, r_1)$$

and for the singlet case

$$\Phi_0^{(0)}(r_1, r_2)\,|_{r_1 \leq r_2} = +\Phi_0^{(0)}(r_2, r_1).$$

These functions on the right-hand sides are the functions actually calculated in the region $r_1 \geq r_2$, the arguments r_1 and r_2 being interchanged.

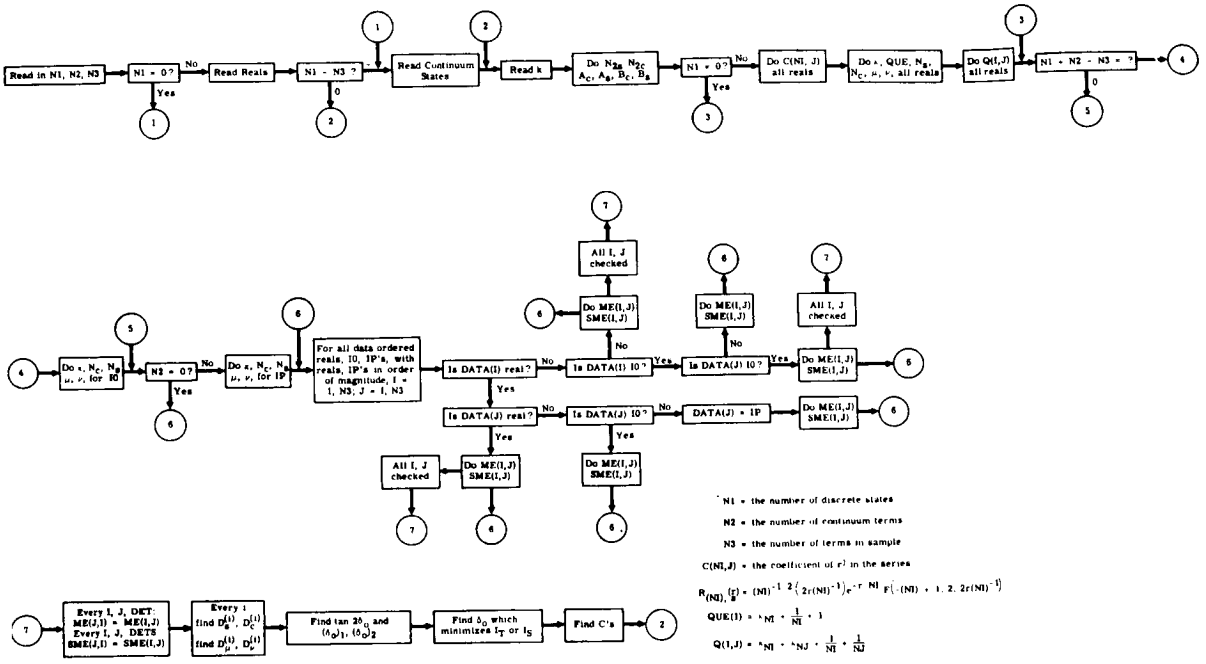


FIGURE 2.—Flow chart for triplet-singlet program.

We have thus evolved formulas for our constants for the triplet and singlet cases. We now turn to a discussion of the programming aspects of these formulas.

For a general idea as to how the programming steps proceed, we refer the reader to the flow chart Fig. 2. Such a program could be modified by the use of sense switch settings to do the triplet part only or the singlet part only. Arrays are properly dimensioned and functions defined when these steps are begun.

That many of the programming steps are trivial is readily implied by the simple algebraic nature of the formulas involved. The continuum states present our most involved nontrivial programming. We describe here in some detail the determination of a singlet matrix element $(\mathrm{SME})_{Ip_1,Ip_2}$. We can use the formula in this form,

$$(\mathrm{SME})_{Ip_1,Ip_2}=2\exp\left[\frac{p_1-p_2}{p_1p_2}\tan^{-1}\frac{p_1-p_2}{\lambda}-\frac{p_1+p_2}{p_1p_2}\tan^{-1}\frac{p_1+p_2}{\lambda}\right]\Big/\lambda^2+(P_1+P_2)^2$$

$$\times\{\cos[\tfrac{1}{2}(a_1-a_2)\ \ln y]+i\sin[\tfrac{1}{2}(a_1-a_2)\ \ln y]\}$$

$$\times\left\{V[2\kappa_{p_1}\kappa_{p_2}+\kappa_{p_2}{}^2+p_2{}^2+(\kappa_{p_1}/\lambda)(\lambda^2+p_1{}^2-p_2{}^2)]\frac{\lambda}{[\lambda^2+(p_1-p_2)^2]}\right.$$

$$\left.-W\left([2\kappa_{p_1}\kappa_{p_2}+\kappa_{p_2}{}^2+p_2{}^2+(\kappa_{p_1}/\lambda)(\lambda^2+p_1{}^2-p_2{}^2)]\frac{2y\lambda[\lambda+i(p_1-p_2)]}{[\lambda^2+(p_1-p_2)^2]^2}-1\right)\right\}.$$

V is the hypergeometric series $F(-ia_2,\ ia_1;\ 1;\ x)$, where $a_j=p_j{}^{-1}$. Hence,

$$V=1+\frac{(-ia_2)(ia_1)x}{1\cdot1!}+\frac{(1-ia_2)(-ia_2)(1+ia_1)(ia_1)x^2}{1\cdot2\cdot2!}+\cdots.$$

W, likewise, is a hypergeometric series $F(1-ia_2,\ 1+ia_1;\ 2;\ x)$.

$$W=1+\frac{(1-ia_2)(1+ia_1)x}{1\cdot2}+\frac{(1-ia_2)(2-ia_2)(1+ia_1)(2+ia_1)x^2}{2!\cdot2\cdot3}+\cdots.$$

We shall call the leading term ($=1$) of both V and W the zeroth term of these respective series. Then, any term $n>0$ of V can be obtained by multiplying the preceding term by the factor

$$[(n-1-1a_2)(n-1+ia_1)x]/n^2,$$

and any term n of W can be obtained by multiplying the preceding term by

$$[(n-ia_2)(n+ia_1)x]/n(n+1).$$

The calculation of this matrix element involves many complex products (V and W are themselves series of complex numbers), so that it was convenient to program a subroutine which would carry out the procedure of complex multiplication, with the possibility of multiplying the complex product by an additional factor EX and dividing by FOFR. The augmented complex product subroutine, which we have named PRODUK, is

```
SUBROUTINE PRODUK (X1, Y1, X2, Y2, EX, FOFR, AA, BB)
AA=(XI*X2-Y1*Y2)*EX/FOFR
BB=(X2*Y1+X1*Y2)*EX/FOFR
RETURN
END (0, 1, 0, 1, 1)
```

Calculation of successive terms were carried out for these series until a continuation resulted in no effect to within the accuracy of the machine, that is, the first eight places of significance. Successive terms together with the partial sums for these series up through convergence are presented in table I for the V and W in $(\mathrm{SME})_{I2,\ I3}$ for $k=0.3$. We see that for series V convergence is reached after summing 15 terms and for series W, further computation was unnecessary after 20 terms. It was seldom necessary to calculate more terms than this for any case. The convergence is the result of the fact that $0<X<1$ for all values of p_1 and p_2 (see below for the definition of X). Also, the fact that the third argument (1 and 2 for V and W) is not a large negative noninteger avoids much convergence difficulty.

To aid in understanding our program of $(\mathrm{SME})_{I_{p1},I_{p2}}$ we present here a dictionary of symbols used. The index I refers to p_1 and the index J refers to p_2.

CAPPA $=(1+p^2-k^2)^{1/2}$

CN1, CNA = real part of factor to be multiplied by previous term of V, W to get current term of V, W.

CN2, CNB = imaginary part of factor to be multiplied by previous term of V, W to get current term of V, W.

$$\mathrm{EM}=\exp\left[\frac{p_1-p_2}{p_1p_2}\tan^{-1}\frac{p_1-p_2}{\lambda}-\frac{p_1+p_2}{p_1p_2}\tan^{-1}\frac{p_1+p_2}{\lambda}\right].$$

$\mathrm{EMALF}=[\lambda^2+(p_1+p_2)^2]^{1/2}$.

$\mathrm{EMGAM}=[\lambda^2+(p_1-p_2)^2]^{1/2}$.

$\mathrm{EN}=p$.

$$\mathrm{FVW1}+i\mathrm{FVW2}=[2\kappa_{p_1}\kappa_{p_2}+\kappa_{p_2}{}^2+p_2{}^2+(\kappa_{p1}/\lambda)(\lambda^2+p_1{}^2-p_2{}^2)]\frac{\lambda V}{[\lambda^2+(p_1-p_2)^2]}$$

$$-\left([2\kappa_{p_1}\kappa_{p_2}+\kappa_{p_2}+p_2{}^2+(\kappa_{p1}/\lambda)(\lambda^2+p_1{}^2-p_2{}^2)]\frac{2y\lambda[\lambda+i(p_1-p_2)]}{[\lambda^2+(p_1-p_2)^2]^2}-1\right)W.$$

TABLE I.—*Successive Terms and Partial Sums for a Sample V and W*[a]

0 IV	Term 1	Term 2	VF1	VF2
1	0.74389441*E*-01	0.0	0.10743894*E* 01	0.0
2	0.96841308*E*-02	0.13834473*E*-02	0.10840736*E* 01	0.13834473*E*-02
3	0.19782332*E*-02	0.44596010*E*-03	0.10860518*E* 01	0.18294074*E*-02
4	0.49964108*E*-03	0.14163055*E*-03	0.10865514*E* 01	0.19710379*E*-02
5	0.14252618*E*-03	0.46825853*E*-04	0.10866939*E* 01	0.20178637*E*-02
6	0.43987560*E*-04	0.16083276*E*-04	0.10867379*E* 01	0.20339470*E*-02
7	0.14344715*E*-04	0.56991363*E*-05	0.10867522*E* 01	0.20396461*E*-02
8	0.48722723*E*-05	0.20708837*E*-05	0.10867570*E* 01	0.20417170*E*-02
9	0.17075203*E*-05	0.76801901*E*-06	0.10867587*E* 01	0.20424850*E*-02
10	0.61345270*E*-06	0.28966723*E*-06	0.10867593*E* 01	0.20427746*E*-02
11	0.22488260*E*-06	0.11080001*E*-06	0.10867596*E* 01	0.20428854*E*-02
12	0.83828040*E*-07	0.42890329*E*-07	0.10867596*E* 01	0.20429283*E*-02
13	0.31691033*E*-07	0.16773310*E*-07	0.10867597*E* 01	0.20429450*E*-02
14	0.12125610*E*-07	0.66179627*E*-08	0.10867597*E* 01	0.20429517*E*-02
15	0.46879206*E*-08	0.26314358*E*-08	0.10867597*E* 01	0.20429543*E*-02

0 IW	Term A	Term B	WF1	WF2
1	0.26036304*E*-00	0.37194721*E*-01	0.12603630*E* 01	0.37194721*E*-01
2	0.79778789*E*-01	0.17984814*E*-01	0.13401418*E* 01	0.55179535*E*-01
3	0.26866236*E*-01	0.76156262*E*-02	0.13670080*E* 01	0.62795161*E*-01
4	0.95797323*E*-02	0.31473454*E*-02	0.13765877*E* 01	0.65942506*E*-01
5	0.35478874*E*-02	0.12972225*E*-02	0.13801356*E* 01	0.67239729*E*-01
6	0.13498287*E*-02	0.53628516*E*-03	0.13814855*E* 01	0.67776013*E*-01
7	0.52397462*E*-03	0.22270727*E*-03	0.13820094*E* 01	0.67998720*E*-01
8	0.20658420*E*-03	0.92918713*E*-04	0.13822160*E* 01	0.68091638*E*-01
9	0.82465024*E*-04	0.38939293*E*-04	0.13822985*E* 01	0.68130577*E*-01
10	0.33253491*E*-04	0.16384048*E*-04	0.13823317*E* 01	0.68146960*E*-01
11	0.13522571*E*-04	0.69187769*E*-05	0.13823452*E* 01	0.68153878*E*-01
12	0.55381978*E*-05	0.29312362*E*-05	0.13823507*E* 01	0.68156809*E*-01
13	0.22820247*E*-05	0.12454923*E*-05	0.13823530*E* 01	0.68158054*E*-01
14	0.94527946*E*-06	0.53060671*E*-06	0.13823540*E* 01	0.68158584*E*-01
15	0.39336933*E*-06	0.22658678*E*-06	0.13823543*E* 01	0.68158811*E*-01
16	0.16436327*E*-06	0.96968222*E*-07	0.13823545*E* 01	0.68158907*E*-01
17	0.68924932*E*-07	0.41578845*E*-07	0.13823546*E* 01	0.68158948*E*-01
18	0.28996781*E*-07	0.17860317*E*-07	0.13823546*E* 01	0.68158966*E*-01
19	0.12234455*E*-07	0.76844777*E*-08	0.13823546*E* 01	0.68158974*E*-01
20	0.51755988*E*-08	0.33112369*E*-08	0.13823546*E* 01	0.68158977*E*-01

[a] These results are for the matrix element $(\mathrm{SME})_{I_2,I_3}$ with $k=0.3$. The symbols IV, IW = the cardinal numbers of the current term of V, W.

Term 1, Term A = real parts of current term of V, W.

Term 2, Term B = imaginary parts of current term of V, W.

VF1, WF1 = real parts of current partial sums for V, W.

VF2, WF2 = imaginary parts of current partial sums for V, W.

IV, IW = the cardinal number of the current term of V, W.

$\mathrm{PCAPFC} = 2\kappa_{p1}\kappa_{p2}+\kappa_{p2}{}^2+p_2{}^2+(\kappa_{p1}/\lambda)(\lambda^2+p_1{}^2-p_2{}^2)$.

$\mathrm{PHI} = [\log_e y]\tfrac{1}{2}(a_1-a_2)$; $\quad a_i = p_i{}^{-1}$.

$\mathrm{SHEDA} = \lambda = \kappa_{p1}+\kappa_{p2}$.

```
76  SHEDA=CAPPA(I)+CAPPA(J)
   0EMALF=SQRTF(SHEDA**2+(EN(I)+EN(J))**
   12)
   0EMGAM=SQRTF(SHEDA**2+(EN(I)-EN(J))**
   12)
      WHY=EMGAM**2/EMALF**2
      PHI=LOGF(WHY)*(1./EN(I)-1./EN(J))/2.
   0EM=EXPF(((EN(I)-EN(J))*ATANF((EN(I)-
   1EN(J))/SHEDA)-(EN(I)+EN(J))*ATANF((E
   2N(I)+EN(J))/SHEDA))/(EN(I)*EN(J)))
   0PCAPFC=2.*CAPPA(I)*CAPPA(J)+CAPPA(J)
   1**2+EN(J)**2+CAPPA(I)*(SHEDA**2+EN(I
   2)**2-EN(J)**2)/SHEDA
      X=4.*EN(I)*EN(J)/EMALF**2

      VF1=1.
      VF2=0.

      TERM1=1.
      TERM2=0.

      IV=1.

1160CALL PRODUK(FLOATF(IV-1),(-1./EN(J))
   1,FLOATF(IV-1),(1./EN(I)),X,FLOATF(IV
   2**2),CN1,CN2)

   0CALL PRODUK(TERM1,TERM2,CN1,CN2,1.,1
   1.,VTERM1,VTERM2)

      TERM1=VTERM1
      TERM2=VTERM2

      VF1=VF1+TERM1
      VF2=VF2+TERM2

   0IF((TERM1**2+TERM2**2)/(VF1**2+VF2**
   12)-1.0E-16)112,114,114

114 IV=IV+1

    GO TO 116
112 WF1=1.
    WF2=0.

    TERMA=1.
    TERMB=0.

    IW=1

1220CALL PRODUK(FLOATF(IW),(-1./EN(J)),F
   1LOATF(IW),(1./EN(I)),X,FLOATF(IW*(IW
   2+1)),CNA,CNB)

   0CALL PRODUK(TERMA,TERMB,CNA,CNB,1.,1
   1.,WTERMA,WTERMB)

    TERMA=WTERMA
    TERMB=WTERMB

    WF1=WF1+TERMA
    WF2=WF2+TERMB

   0IF((TERMA**2+TERMB**2)/(WF1**2+WF2**
   12)-1.0E-16)118,120,120

120 IW=IW+1

    GO TO 122
1180CALL PRODUK(PCAPFC*2.*WHY*SHEDA**2/E
   1MGAM**4-1.,PCAPFC*2.*WHY*SHEDA*(EN(I
   2)-EN(J))/EMGAM**4,WF1,WF2,1.,1.,TYW0
   3L1,TYWOL2)

   0FVW1=VF1*PCAPFC*SHEDA/EMGAM**2-TYWOL
   11
   0FVW2=VF2*PCAPFC*SHEDA/EMGAM**2-TYWOL
   12

   0CALL PRODUK(COSF(PHI),SINF(PHI),FVW1
   1,FVW2,2.*EM,EMALF**2,SME1,SME2)

    SME(I,J)=SME1
```

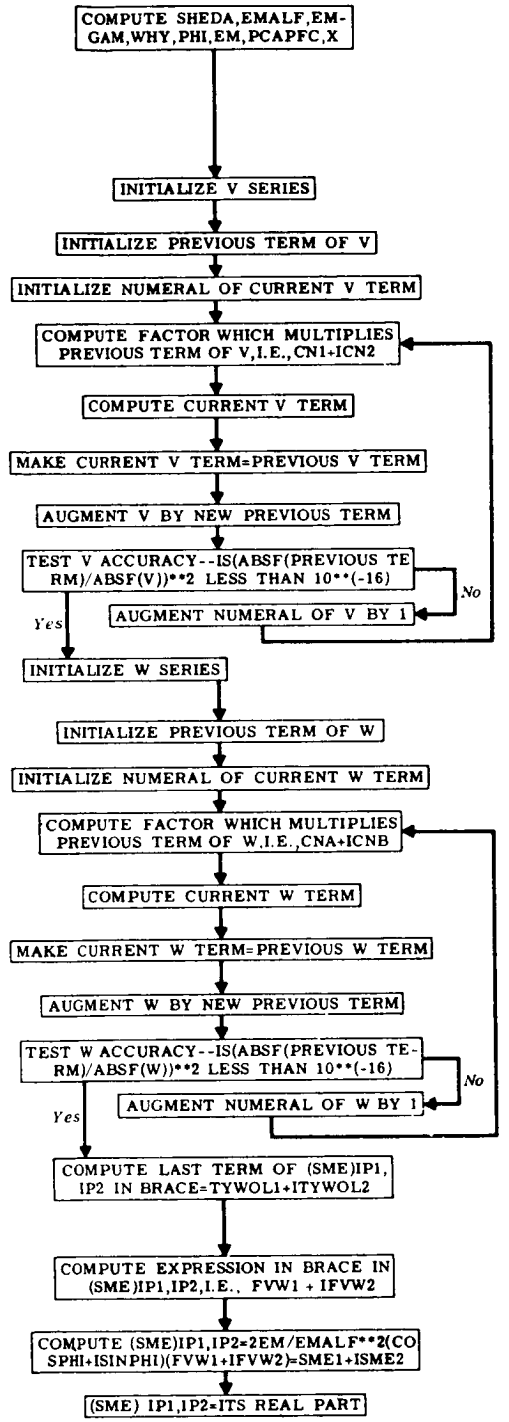


FIGURE 3.—Flow chart and program for $(SME)_{Ip_1,Ip_2}$. See text for definitions of symbols.

$\mathrm{SME1}+i\mathrm{SME2}=(\mathrm{SME})_{Ip_1,Ip_2}$.

TERM1, TERMA = real parts of previous term of V, W.

TERM2, TERMB = imaginary parts of previous term of V, W.

$$\mathrm{TYWOL1}+i\mathrm{TYWOL2}=\left(\left[2\kappa_{p_1}\kappa_{p_2}+\kappa_{p_2}{}^2+p_2{}^2+(\kappa_{p_1}/\lambda)\,(\lambda^2+p_1{}^2-p_2{}^2)\right]\frac{2y\lambda[\lambda+i(p_1-p_2)]}{[\lambda^2+(p_1-p_2)^2]^2}-1\right)W.$$

VF1, WF1 = real parts of hypergeometric series V, W.

VF2, WF2 = imaginary parts of hypergeometric series V, W.

VTERM1, WTERMA = real parts of current term of V, W.

VTERM2, WTERMB = imaginary parts of current term of V, W.

$\mathrm{WHY}=\mathrm{Y}=[\lambda^2+(p_1-p_2)^2]/[\lambda^2+(p_1+p_2)^2]$.

$\mathrm{X}=4p_1p_2/[\lambda^2+(p_1+p_2)^2]$.

We present in Fig. 3 a faithful excerpt from our operational singlet program which calculates the the matrix element $(\mathrm{SME})_{Ip_1,Ip_2}$. Included along side these Fortran steps is a schematic diagram of the logic paths and decisions employed to serve as a reference for quicker understanding when reading these statements.

The discrete-discrete states and discrete-continuum states are explicitly real in their analytic form. The formulas for the continuum-continuum cases seem to be complex, but we hereby assert and show by numerical example the reality of the matrix elements for these states also. (An analytic demonstration is given in the Appendix II.)

1. Example

The analytic form of the triplet continuum-continuum matrix element may be written as follows

$$(\mathrm{ME})_{Ip_1,Ip_2}=2\lambda\exp\left[\frac{p_1-p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1-p_2}{\lambda}\right)-\frac{p_1+p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1+p_2}{\lambda}\right)\right]\Big/\,[\lambda^2+(p_1-p_2)^2][\lambda^2+(p_1+p_2)^2]$$

$$\times\{\cos[\tfrac{1}{2}(a_1-a_2)\,\ln y]+i\sin[\tfrac{1}{2}(a_1-a_2)\,\ln y]\}\left\{V-\frac{2y[\lambda+i(p_1-p_2)]}{\lambda^2+(p_1-p_2)^2}W\right\}. \qquad (36)$$

V and W are hypergeometric series discussed above. This case we demonstrate is calculated for these values of the parameters

$$Ip_1=I_2,$$

$$Ip_2=I_3,$$

$$k=0.3.$$

$$\kappa_{Ip_i}=(1+p_i{}^2-k^2)^{1/2}; \qquad \kappa_{I2}=2.2158520; \qquad \kappa_{I3}=3.1480152.$$

$$\lambda=\kappa_{Ip_1}+\kappa_{Ip_2}=5.3638672.$$

$$a_i=(1/p_i); \qquad a_1=0.5; \qquad a_2=0.33333333.$$

$$y=[\lambda^2+(p_1-p_2)^2]/[\lambda^2+(p_1+p_2)^2]=0.55366334.$$

$$V=1.0867597+i0.0020429543.$$

$$W=1.3823546+i0.068158977.$$

Then

$$2y/[\lambda^2+(p_1-p_2)^2]=0.037194721.$$

$$\{2y[\lambda+i(p_1-p_2)]/\lambda^2+(p_1-p_2)^2\}=0.19950754-i0.037194721.$$

$$\{2y[\lambda+i(p_1-p_2)]/[\lambda^2+(p_1-p_2)^2]\}W=0.27832532-i0.037818064.$$

$$\left\{V-\frac{2y[\lambda+i(p_1-p_2)]}{\lambda^2+(p_1-p_2)^2}W\right\}=0.80843438+i0.039861018.$$

$$\{\cos[\tfrac{1}{2}(a_1-a_2)\ \ln y]+i\sin[\tfrac{1}{2}(a_1-a_2)\ \ln y]\}=0.99878665-i0.049246611.$$

$$\{\cos[\tfrac{1}{2}(a_1-a_2)\ \ln y]+i\sin[\tfrac{1}{2}(a_1-a_2)\ \ln y]\}\left\{V-\frac{2y[\lambda+i(p_1-p_2)]}{\lambda^2+(p_1-p_2)^2}W\right\}=0.80941649+i0.$$

$$\exp\left[\frac{p_1-p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1-p_2}{\lambda}\right)-\frac{p_1+p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1+p_2}{\lambda}\right)\right]=0.55182012.$$

$$2\lambda\exp\left[\frac{p_1-p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1-p_2}{\lambda}\right)-\frac{p_1+p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1+p_2}{\lambda}\right)\right]\Big/[\lambda^2+(p_1-p_2)^2][\lambda^2+(p_1+p_2)^2]=0.0036979614,$$

and finally,

$$(\mathrm{ME})_{I_{p1},I_{p2}}=0.0036979614\times0.80941649=0.0029931909,$$

a real result, the imaginary part vanishing to all eight figures.

This complete cancellation is impressive but coincidental, for, in general, in computing where eight significance figures are retained, the accuracy of the last place(s) is questionable. The singlet continuum-continuum matrix element for the same parameters provides a more realistic example in that the imaginary part is zero only to the accuracy of the real part.

We present the analytic form of the singlet continuum-continuum matrix element in the following form since we have already obtained the triplet numerical answer.

2. Example

$$(\mathrm{SME})_{I_{p1},I_{p2}}=[2\kappa_{p1}\kappa_{p2}+\kappa_{p_2}{}^2+p_2{}^2+(\kappa_{p1}/\lambda)(\lambda^2+p_1{}^2-p_2{}^2)](\mathrm{ME})_{I_{p1},I_{p2}}$$
$$+2\exp\left[\frac{p_1-p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1-p_2}{\lambda}\right)-\frac{p_1+p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1+p_2}{\lambda}\right)\right]\Big/\lambda^2+(p_1+p_2)^2$$
$$\times\{\cos[\tfrac{1}{2}(a_1-a_2)\ \ln y]+i\sin[\tfrac{1}{2}(a_1-a_2)\ \ln y]\}W$$

and here,

$$[2\kappa_{p1}\kappa_{p2}+\kappa_{p_2}{}^2+p_2{}^2+(\kappa_{p1}/\lambda)(\lambda^2+p_1{}^2-p_2{}^2)](\mathrm{ME})_{I_{p1},I_{p2}}=42.681072\times0.0029931909=0.12775260.$$

$$2\exp\left[\frac{p_1-p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1-p_2}{\lambda}\right)-\frac{p_1+p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1+p_2}{\lambda}\right)\right]\Big/\lambda^2+(p_1+p_2)^2=0.020524796.$$

$$\{\cos[\tfrac{1}{2}(a_1-a_2)\ \ln y]+i\sin[\tfrac{1}{2}(a_1-a_2)\ \ln y]\}W$$
$$=(0.99878665-i0.049246611)\times(1.3823546+i0.068158977)$$
$$=1.3840339-i0.3\times10^{-8}.$$

$$2\exp\left[\frac{p_1-p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1-p_2}{\lambda}\right)-\frac{p_1+p_2}{p_1p_2}\tan^{-1}\left(\frac{p_1+p_2}{\lambda}\right)\right]\Big/\lambda^2+(p_1+p_2)^2$$
$$\times\{\cos[\tfrac{1}{2}(a_1-a_2)\ \ln y]+i\sin[\tfrac{1}{2}(a_1-a_2)\ \ln y]\}W=0.020524796(1.3840339-i0.3\times10^{-8})$$
$$=0.028407013-i0.61574388\times10^{-10}.$$

then

$$(\mathrm{SME})_{I_{p1},I_{p2}}=0.12775260+0.028407013-i0.61574388\times10^{-10}$$
$$=0.15615961-i0.61574388\times10^{-10}.$$

The first seven places of significance in the imaginary part cancel in the complex product

$$\{\cos[\tfrac{1}{2}(a_1-a_2)\ \ln y]+i\sin[\tfrac{1}{2}(a_1-a_2)\ \ln y]\}W,$$

above. Only the eighth place remained to give the nonzero imaginary part in the final answer, the accuracy of which place we have already laid open to question.

We could check the correctness of the analytic expressions for integrals involving continuum Coulomb wave functions by numerically integrating some of the integrals. The numerical determination of the continuum function $u_p(r)$ is required in the next section. There one will find a discussion of the method by which these functions were obtained. The remaining factors in each integral are trivially obtained. Table II contains a comparison of values from numerical integration with answers from evaluating explicit formulas. The comparison shows rather cogently that the analytic formula, as well as our technique for evaluating them, are correct.

To acquaint the reader with some of the difficulties and limitations with which we were plagued in our probes for numerical answers we detail here the evaluation of two determinants.

We have chosen at random as one example a third-order determinant composed of the triplet matrix elements for the discrete states

$$n=2,\ 3,\ 4$$

and

$$k=0.05.$$

3. Example

Evaluate the determinant

$$\begin{vmatrix} 0.014932179 & 0.0075449045 & 0.0047813990 \\ 0.0075449045 & 0.0038303733 & 0.0024309322 \\ 0.0047813990 & 0.0024309322 & 0.0015434798 \end{vmatrix}.$$

We expand this determinant by summing the products of the elements of the first row by their respective cofactors,

$$0.014932179(0.0038303733\times0.0015434798-0.0024309322\times0.0024309322)$$
$$=0.014932179(0.59121038\times10^{-5}-0.59094314\times10^{-5})$$
$$=0.014932179(0.26724\times10^{-8})=0.39904755\times10^{-10}.$$

$$-0.0075449045(0.0075449045\times0.0015434798-0.0024309322\times0.0047813990)$$
$$=-0.0075449045(0.11645408\times10^{-4}-0.11623257\times10^{-4})$$
$$=-0.0075449045(0.22151\times10^{-7})=-0.16712\mathit{718}\times10^{-9}$$

$$0.0047813990(0.0075449045\times0.0024309322-0.0038303733\times0.0047813990)$$
$$=0.0047813990(0.18341151\times10^{-4}-0.18314543\times10^{-4})$$
$$=0.0047813990(0.26608\times10^{-7})=0.12722\mathit{346}\times10^{-9}$$

Then

$$0.399047\mathit{55}\times10^{-10}-0.16712\mathit{718}\times10^{-9}+0.12722\mathit{346}\times10^{-9}=0.\mathit{1035}\times10^{-14}.$$

TABLE II.—*Numerical Evaluation of Analytic Formulas Versus Numerical Integration of Integrals for Certain Matrix Elements*

DATA (i)	DATA (j)	Analytic formulas evaluation	Numerical integration
		$(ME)_{ij}$	
3	I_0	0.94464160×10^{-2}	0.94463×10^{-2}
4	I_2	0.21539854×10^{-2}	0.215391×10^{-2}
I_0	I_3	0.45762914×10^{-2}	0.45760×10^{-2}
I_2	I_3	0.29639473×10^{-2}	0.29636×10^{-2}
		$\mathfrak{N}_{si}$	
I_0		0.23023649×10^{-2}	0.230233×10^{-2}
I_2		0.37946834×10^{-2}	0.37945×10^{-2}
		$\mathfrak{N}_{ci}$	
I_1		0.55343236×10^{-1}	0.66345×10^{-1}
I_3		0.12038466×10^{-1}	0.12041×10^{-1}

We have italicized the figures that are not significant as a result of subtraction. Thus one sees, in this computation, that all 8 places were lost, which is all the significance the machine affords. When we check this result against the answer the machine gives, we get agreement only in order of magnitude and sign because what appears to be remaining significant figures have no meaning. The value of this determinant as given by the machine is 0.37339169×10^{-14}.

Our next example is $\mathfrak{D}_c^{(2)}$ for the determinant above, the second row having been replaced by $\mathfrak{N}_{ci}$.

4. Example

Evaluate the determinant

$$\begin{vmatrix} 0.014932179 & 0.0075449045 & 0.0047813990 \\ 0.078255467 & 0.038279708 & 0.024017526 \\ 0.0047813990 & 0.0024309322 & 0.0015434798 \end{vmatrix}.$$

Expanding as in the previous example we get

$$0.10437\mathit{280}\times10^{-7}-0.17334\mathit{368}\times10^{-8}$$
$$-0.87051\mathit{106}\times10^{-8}-0.12\mathit{674}\times10^{-11}.$$

The loss in significant figures is not as complete as in the first example. Two questionable places were salvaged which compare surprisingly favorably with the machine answer. The machine gave $-0.12508288\times10^{-11}$ for the value of this determinant.

This tremendous loss of accuracy plus addition of negative exponents in multiplication while evaluating our determinants emphasized the fact heretofore expressed of our being limited to finite samplings of states. Regardless of terms chosen and the k being investigated, determinants of order 10 and above were usually given a value of zero by the machine because of the small magnitude of the result. See the table III for a typical illustration of how the value of these determinants decrease in magnitude with increasing number of terms soon vanishing to the accuracy of the machine.

As was pointed out earlier in numerical display,

TABLE III.—*Vanishing of Determinants with Increasing Order ($k = 0.05$)*

States	Determinants	I_T	δ_0
2	0.14932180×10^{-1}	0.25772093×10^{-3}	0.30331454×10
2, 3	0.27025105×10^{-6}	0.38947910×10^{-5}	0.30244533×10
2, 3, 4	0.37339169×10^{-14}	0.24750829×10^{-4}	0.30289188×10
2, 3, 4, 5	$-0.25949299\times10^{-24}$	0.24039984×10^{-1}	0.27652040×10^{-1}
2, 3, 4, 5, I_0	$-0.41640429\times10^{-32}$	0.24514973×10^{-1}	0.28817544×10^{-1}
2, 3, 4, 5, I_0, I_1	$-0.15898579\times10^{-36}$	0.22766215×10^{-1}	0.25508314×10^{-1}
2, 3, 4, 5, I_0, I_1, I_2	0	0.18691529×10^{-2}	0.30666550×10
2, 3, 4, 5, I_0, I_1, I_2, I_3	0	0.18691529×10^{-2}	0.30666550×10

we have lost all accuracy in the value of our determinant by the time we reach the three states 2, 3, 4. It is easy to see that if there were no loss in accuracy the value of the integral I_T must decrease as you add terms to $\Phi_0^{(0)}$ without deleting what is already there. Referring to table III, we see that this monotonic decrease takes place only until 2, 3, 4. There and beyond we can be sure that no significant figures remain in I_T. We do not wish to imply that there is any one-to-one correspondence between the number of significant figures in I_T and the number of significant figures in the determinants. And particularly we wish no such implication between I_T and the phase shift. In fact, we found the greatest uniformity in the phase shifts from several different expansions was obtained, when all of these expansions gave a negative value for I_T or I_S (in which case one could be sure no significant figures survived in I_T or I_S).

In our operational programs for the triplet and singlet phase shifts and constants, the FORTRAN subroutine, RWDET, SHARE* 635, was used to evaluate our determinants.

We have already indicated that our formula for tan $2\delta_0$ yields two values of δ_0 which we call $\delta_{0_{1,2}}$. For each δ_0 there is an associated array of C_i's. Working then with these two sets of constants, we obtain two answers for our square deviation.

We examine next the results of a sample calculation (see table IV) of this quantity for the triplet case,

$$I_T=\int_0^\infty \sin^2(kr+\delta_0)\,R_{1s}{}^2(r)\,dr + 2\sum_i C_iY_i+\sum_i C_i{}^2(\mathrm{ME})_{ii} + 2\sum_{i>j}\sum C_iC_j(\mathrm{ME})_{ij},$$

$$\int_0^\infty \sin^2(kr+\delta_0)\,R_{1s}{}^2(r)\,dr = \tfrac{1}{2}\{1-\mathfrak{N}_{2c}\cos 2\delta_0+\mathfrak{N}_{2s}\sin 2\delta_0\},$$

a function of δ_0 and k only. We present the two resulting partial sums as the indices run through their ranges of values. The result of smallest magnitude given finally as the value of EMS: we take as the value of I and the δ_{0_i} that gave the result we designate as δ_0.

In this example I_T equals -0.16530976×10^{-7}. And therefore $\delta_0=0.31182027\times10$. With the correct δ_0 identified, the associated set of C_i's could be uniquely determined and are the ones given in table IV. One can also observe there that all the significant figures in I_T cancel. This, then, is the source of the negative values of I_T, which is theoretically positive definite. This implies therefore that the result finally assigned to I_T is probably meaningful only in order of magnitude of the absolute value.

In summary, the operational program used to solve the zeroth order problem determines tan $2\delta_0$ from which is obtained a pair of δ_0's and their associated sets of C_i's. These two sets of constants yield two values of the total square deviation I_T or I_S; the program chooses the δ_0 that gives the smaller total square deviation as the phase shift for the particular energy investigated. The corresponding set of C_i's which completes the specification of $\Phi_0^{(0)}$ are then printed out.

Determining a set of terms to yield a minimum total is to a great extent a matter of trial and error, but the ability to do so was greatly enhanced by practice. We found that for negative totals, the δ_0's are usually about as correct as the program can yield in that this is a zero value of I_T or I_S to the accuracy of the machine, as previously illustrated. In the absence of a negative total square deviation, we looked for the smallest values of I_T or I_S attainable, which, when these values are suitably small, give rise to a nearly constant set of δ_0's. In table V we have presented δ_0 and the corresponding I_S for various expansions. The good expansions are on the left. All the I_S for a given k are about the same order of magnitude, and the δ_0 are essentially independent of which good expansion we used. In each case the good expansions were obtained from a much larger sample of states than the number of poor expansions given on the right. These samples, nevertheless, show the wild variation in δ_0 when the minimum I_S was not obtained. Although the

* SHARE is a voluntary organization of IBM 704, 709, and 7090 lessors and owners constituted to distribute routines of general interest. It is assumed in various places of this article that the reader is familiar with the FORTRAN system of programming.

TABLE IV.—*Determining a Minimum* I_T*

States: 2, 3, I_0; $k = 0.01$; $\tan 2\delta_0 = -0.046814054$

$\delta_{0_1} = 3.1182027$

I	J	$\int_0^\infty \sin^2(kr + \delta_0) R_{1s}^2(r)\, dr$	$C_i^2(\mathrm{ME})_{ii}$	$2C_iY_i$	$2C_iC_j(\mathrm{ME})_{ij}$	I_{T_1}
		$0.14535630 \times 10^{-3}$				$0.14535630 \times 10^{-3}$
1			$0.91008559 \times 10^{-2}$			$0.92462122 \times 10^{-2}$
				$-0.22028068 \times 10^{-2}$		$0.70434052 \times 10^{-2}$
2			$0.17929179 \times 10^{-1}$			$0.24972585 \times 10^{-1}$
				$0.30211914 \times 10^{-2}$		$0.27993776 \times 10^{-1}$
	1				$-0.25487592 \times 10^{-1}$	$0.20561937 \times 10^{-2}$
3			$0.24987741 \times 10^{-2}$			$0.50049677 \times 10^{-2}$
				$-0.11074711 \times 10^{-2}$		$0.38974967 \times 10^{-2}$
	1				$0.94745301 \times 10^{-2}$	$0.13372026 \times 10^{-1}$
	2				$-0.13372043 \times 10^{-1}$	$-0.16530976 \times 10^{-7}$

$\therefore \delta_j = 3.1182027$

States:	2	3	I_0
C_i:	-0.78149776	2.1654110	-0.32740131

$\delta_{0_2} = 1.5474064$

I	J	$\int_0^\infty \sin^2(kr + \delta_0) R_{1s}^2(r)\, dr$	$C_i^2(\mathrm{ME})_{ii}$	$2C_iY_i$	$2C_iC_j(\mathrm{ME})_{ij}$	I_{T_2}
		0.99985464				0.99985464
1			0.66812128×10^{2}			0.67811982×10^{2}
				$-0.10466429 \times 10^{2}$		0.57345552×10^{2}
2			0.15130557×10^{3}			0.20865113×10^{3}
				0.15212064×10^{2}		0.22386319×10^{3}
	1				$-0.20061456 \times 10^{3}$	0.23248637×10^{2}
3			0.22729574×10^{2}			0.45978212×10^{2}
				-0.57429675×10		0.40235244×10^{2}
	1				0.77424136×10^{2}	0.11765938×10^{3}
	2				$-0.11715941 \times 10^{3}$	0.49996948

* Tabulation of addends and partial sums; designation of δ_j and associated C_i's.

TABLE V.—*δ_0 and I_S for Various Expansions of $\Phi_0^{(0)}$*

k	Good expansions			Poor expansions		
	I_S	δ_0	States	I_S	δ_0	States
0.01	-0.3725×10^{-8}	3.06376	3, 4	0.4263×10^{-3}	1.87084	3, 4, I_0
	-0.1982×10^{-7}	3.06356	3, I_0, I_3	0.2645×10^{-3}	3.13968	2, 3, 4, I_1
	-0.1863×10^{-7}	3.06372	4, I_0	0.5538×10^{-3}	0.0192341	2, 3, 4, I_0, I_2, I_3, I_4
	0.1088×10^{-7}	3.06353	2, I_0, I_1, I_2, I_3	0.1018×10^{-2}	0.0589352	3, 4, I_0, $I_{0.5}$
0.1	0.7741×10^{-6}	2.42070	5, I_0, I_5	0.1867	2.20460	I_3
	0.7833×10^{-6}	2.41736	4, I_0, I_2, I_3, I_4	0.2213×10^{-1}	3.06826	2, 4, 5, I_0, I_3
	0.9183×10^{-6}	2.41822	4, I_0, I_3	0.7176×10^{-1}	0.473490	2, 3, 4, 5, I_2
	0.9344×10^{-6}	2.41772	4, I_0, I_4, I_8, I_{12}, I_{16}	0.4385×10^{-1}	1.99190	5, I_0, I_5, I_6, I_7, I_8, I_9
0.75	0.1170×10^{-4}	0.756407	2, 3, I_0, $I_{0.5}$, $I_{1.0}$, $I_{1.5}$, $I_{2.0}$	0.4353×10^{-1}	0.539717	2, 3, 4, I_0, $I_{0.25}$
	0.1094×10^{-4}	0.753542	2, 3, 4, $I_{0.5}$, $I_{1.0}$, $I_{1.5}$, $I_{2.0}$, $I_{2.5}$	0.2796×10^{-1}	0.878779	3, 5, I_0, $I_{0.5}$
	0.5165×10^{-5}	0.755573	2, 3, 4, $I_{0.75}$, $I_{1.5}$, $I_{2.25}$, $I_{3.0}$, $I_{3.75}$, $I_{4.5}$	0.5196×10^{-1}	0.908830	3, 4, $I_{0.75}$
	0.1314×10^{-4}	0.756173	2, 4, 5, $I_{0.75}$, $I_{1.5}$, $I_{2.25}$, $I_{3.0}$, $I_{3.75}$	0.1587	0.358244	2, 3, 4, 5, I_0, $I_{0.25}$

magnitude of the minimum I_S is seen to go up with increasing k, it probably represents almost a maximum cancellation of significant figures in all the terms which comprise I_S. Thus, the rise in magnitude of I_S reflects a rise in the magnitude of the individual terms with k.

Referring to Eq. (26), having evaluated a δ_0 and the associated C_i's, we are enabled to evaluate $\Phi_0^{(0)}$ in the region $r_2 \leq r_1$. In Fig. 4 we present results for two different expansions of $\Phi_0^{(0)}$. The topmost set of numbers were obtained by numerical integration.*

Aside from the fact that the numerical solution requires a finite mesh size ($\Delta r = 0.1$ in this case), the numerical solution is not exact because the outer boundary was taken at $r_1 = 10$. We think that this solution is nevertheless accurate to within one or two in the third figures in the region presented. (There may also be a slight overall renormalization of the integrated results.) The procedure we use thus seems to yield very satisfactory results. The numerical evaluation (and accuracy) of $\Phi_0^{(0)}$ is necessary in obtaining the higher order corrections (Section IV).

The zeroth-order problem is a very simplified type of many-body scattering problem. Nevertheless the results are very respectable approximations of the final answers (cf. Section V). In addition, this idealized problem contains implications for the low energy scattering of electrons from atoms in general (cf. Temkin, 1962) which are perhaps more important than its specific applicability to hydrogen.

The technique of expanding about exact separate solutions and minimizing with respect to the nonseparable boundary condition is new to the electron scattering problem[2] as far as we know. It has the virtue of preserving both the physically meaningful and mathematically rigorous aspects of the solution, the latter property having heretofore largely been absent from treatments of this type of problem. It also has a practical advantage: on retaining only two terms in the expansion, which leads to a comparatively easy hand calculation, one gets almost as good results (Temkin, 1960) as the much more difficult exchange approximation (cf. Section V). The carrying of the program in full, however, has been a distinctly nontrivial task of which, it is hoped, this section has given some idea. It clearly would have been impossible without the availability of large scale computing machines.

Nevertheless, the phenomenon of the cancellation of significant figures in the determinants cannot be classed as completely satisfactory. We are fortunate, in fact, that quantities in which we are interested are fairly insensitive to this cancellation.

CALCULATION OF THE HIGHER ORDER TERMS

It is clear from the Eqs. (10) that the higher order functions Φ_l that enter in the integrands of the integrals of the basic relation Eq. (23) can only be calculated iteratively. We shall introduce a perturbation theory that effectively formilizes an iteration procedure. No claim is made for the uniqueness of this perturbation theory, and in fact the character of the present treatment is somewhat provisional.

In order to get an idea of what the perturba-

* The numerical integration was carried out by E. Sullivan in consultation with Mr. W. Cahill, head of the mathematics branch of this division. The integration of Eq. (17) is obviously a modest enterprise compared to Eq. (1), nevertheless a high order of competence was required. In particular the ordinary point-by-point relaxation technique would not converge because of the variability in sign of $E+2r_2^{-1}$. Sullivan and Cahill worked very hard on a block relaxation technique until they showed that this too would not converge. Sullivan finally successfully introduced a noniterative technique, in which in essence each grid point is directly related to the boundary values. This experience, it seems to us, constitutes a very good example of the difficulty that even a two-dimensional equation may present for numerical solution. In addition, it indicates rather strongly that the full three-dimensional equation will not be soluble by an iterative technique.

Note added in proof: The numerical integration mentioned above has now been completed in the triplet case (Temkin and Sullivan, to be published). In addition, the zeroth order problem has been extended to double precision. The main results of this investigation are: (1) the double precision program affords greatly increased accuracy ($\sim$2 more significant figures) over the single precision program, and the resultant $\Phi_0^{(0)}$ are believed to be more reliable than the numerically integrated $\Phi_0^{(0)}$ for the purposes of integrating the higher order equations; (2) the final results generally substantiate larger phase shifts than Schwartz's, although they are closer to Schwartz's than to the present results. Specifically, the triplet scattering length is found to be $a_t = 1.7683$. If the numerical integrations have the requisite accuracy, this can be taken as a reliable upper bound on the scattering length.

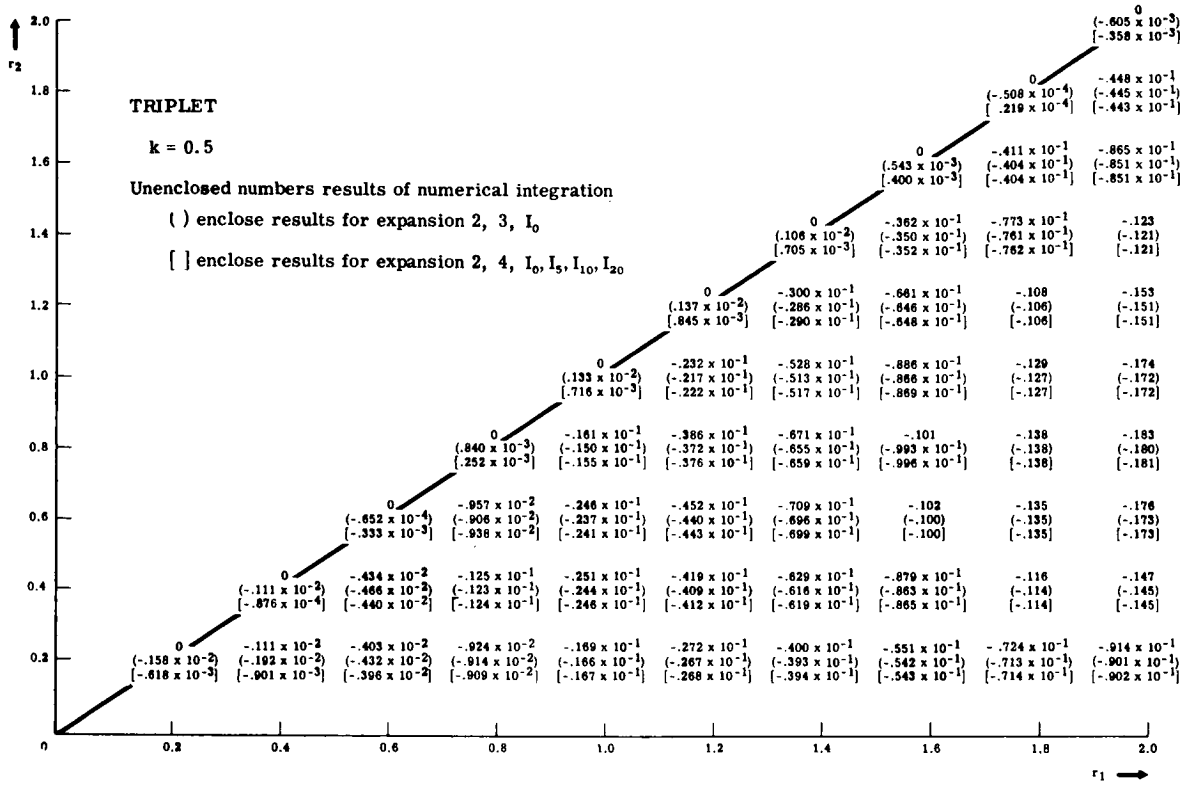


FIGURE 4.—Numerical values of $\phi_0^{(0)}$ for $k=0.5$ triplet case.

tion theory consists, let us write the first several equations (10) more explicitly:

$$\left\{\Delta_{12}+\frac{2}{r_2}+E\right\}\Phi_0(r_1, r_2)=\sum_{m=1}^{\infty}\frac{2}{(2m+1)^{1/2}}\frac{r_2{}^m}{r_1{}^{m+1}}\Phi_m(r_1, r_2) \tag{37a}$$

$$\left\{\Delta_{12}-2\left(\frac{1}{r_1{}^2}+\frac{1}{r_2{}^2}\right)+\frac{2}{r_2}-\frac{4}{5}\frac{r_2{}^2}{r_1{}^3}+E\right\}\Phi_1=\frac{2}{\sqrt{3}}\frac{r_2}{r_1{}^2}\Phi_0+(\quad)\Phi_2+\cdots \tag{37b}$$

$$\left\{\Delta_{12}-6\left(\frac{1}{r_1{}^2}+\frac{1}{r_2{}^2}\right)+\frac{2}{r_2}-\frac{4r_2{}^2}{7r_1{}^3}-\frac{4}{7}\frac{r_2{}^4}{r_1{}^5}+E\right\}\Phi_2=\frac{2}{(5)^{1/2}}\frac{r_2{}^2}{r_1{}^3}\Phi_0+\cdots. \tag{37c}$$

The perturbation theory itself is based on the assumption that the dominant contribution to the integrals comes from the adiabatic region. As we shall see, this assumption is justified at the lowest energies. The fact that the adiabatic form of Φ_l, Eq. (15), is known is of paramount importance in this connection.

The quantity which enters the equations and whose size conveniently measures the adiabaticity of a region of configuration space is $r_2{}^n/r_1{}^{n+1}$. We shall associate with this quantity an order of magnitude $\lambda^{n/2}$, corresponding to the fact that for all but an amount of measure zero of our configuration space $r_2\leq r_1$, the larger n the smaller this quantity. We also expand Φ_l in a somewhat arbitrary, but not unreasonable way,

$$\Phi_l=\sum_{j=0}^{\infty}\lambda^{j+1/2}l\Phi_l{}^{(j)} \tag{38}$$

This says two things: (1) That in the adiabatic region higher multipoles get successively smaller, and

(2) that the behavior of any multipole in the nonadiabatic region can be expanded about its behavior in the adiabatic region. In order to be useful, Eq. (38) should *either* converge rapidly as a function of l, which means that the behavior in the nonadiabatic region should not differ greatly from the form applicable to the adiabatic region, *or* the contribution from the nonadiabatic region itself should be small compared to the contribution from the adiabatic region. In a generally overlapping fashion these assumptions apply to the lower multipoles and energies.

The substitution of Eq. (38) into the set of Eqs. (37) with the use of the order of magnitude associated with r_2^n/r_1^{n+1} yields a set of equations characterized by different powers of $\lambda^{1/2}$. The ones that we attempt to take account of in this calculation are

$$\lambda^0: \qquad [\Delta_{12}+(2/r_2)+E]\Phi_0^{(0)}=0 \tag{39a}$$

$$\lambda^{1/2}: \qquad \left[\Delta_{12}-2\left(\frac{1}{r_1^2}+\frac{1}{r_2^2}\right)+\frac{2}{r_2}+E\right]\Phi_1^{(0)}=\frac{2}{\sqrt{3}}\frac{r_2}{r_1^2}\Phi_0^{(0)} \tag{39b}$$

$$\lambda: \qquad \left(\Delta_{12}+\frac{2}{r_2}+E\right)\Phi_0^{(1)}=\frac{2}{\sqrt{3}}\frac{r_2}{r_1^2}\Phi_1^{(0)} \tag{39c}$$

$$\left[\Delta_{12}-6\left(\frac{1}{r_1^2}+\frac{1}{r_2^2}\right)+\frac{2}{r_2}+E\right]\Phi_2^{(0)}=\frac{2}{(5)^{1/2}}\frac{r_2^2}{r_1^3}\Phi_0^{(0)} \tag{39d}$$

$$\lambda^{3/2}: \quad \left[\Delta_{12}-2\left(\frac{1}{r_1^2}+\frac{1}{r_2^2}\right)+\frac{2}{r_2}+E\right]\Phi_1^{(1)}-\frac{4}{5}\frac{r_2^2}{r_1^3}\Phi_1^{(0)}=\frac{2}{\sqrt{3}}\frac{r_2}{r_1^2}\Phi_0^{(1)}+\frac{4}{(15)^{1/2}}\frac{r_2}{r_1^2}\Phi_2^{(0)} \tag{39e}$$

The first equation (39a) is the zeroth-order equation (17), the physical significance and solution of which we have discussed in Section III. Comparing the remaining equations of (39) with the Eqs. (37), one sees that terms which are omitted in the former tend to vanish or becomes small, in comparison with the terms that are retained, in the adiabatic region. This, of course, is built into the perturbation theory.

The inserting of the same perturbation expansions into the basic relation, Eq. (23) yields an expansion in integral powers of λ,

$$\sin(\delta-\delta_0)=-k^{-1}\sum_{\nu=1}^{\infty}\lambda^\nu\left\{\sum_{\substack{m+\mu=\nu \\ (\substack{m\geq 1 \\ \mu\geq 0})}}\frac{2}{(2m+1)^{1/2}}\int_0^\infty\int_0^{r_1}\Phi_0^{(0)}(r_2^m/r_1^{m+1})\Phi_m^{(\mu)}\,dr_1\,dr_2\right\}. \tag{40}$$

The quantity in curly brackets is expected to have the order of magnitude λ^ν, so that ν characterizes the order of our perturbation theory. (However, λ is to be set equal to one for numerical evaluation.) The first-order correction is

$$\Delta\delta_0=-k^{-1}\frac{3}{k\sqrt{3}}\int_0^\infty\int_0^{r_1}\Phi_0^{(0)}(r_2/r_1^2)\Phi_1^{(0)}\,dr_1\,dr_2. \tag{41}$$

The second-order correction consists of two terms,

$$\Delta^2\delta_0^{(1)}=-k^{-1}\frac{2}{\sqrt{3}}\int_0^\infty\int_0^{r_1}\Phi_0^{(0)}(r_2/r_1^2)\Phi_1^{(1)}\,dr_1\,dr_2 \tag{42}$$

and

$$\Delta^2\delta_0^{(2)}=-k^{-1}\frac{2}{(5)^{1/2}}\int_0^\infty\int_0^{r_1}\Phi_0^{(0)}(r_2^2/r_1^3)\Phi_2^{(0)}\,dr_1\,dr_2. \tag{43}$$

The quantities $\Delta\delta_0$ and $\Delta^2\delta_0^{(2)}$ are the obvious perturbation theoretic counterparts of the first two terms

(dipole and quadrupole) of the basic relation. But in perturbation theory there is a term additional to the quadrupole term which is expected to be second order in magnitude, $\Delta^2\delta_0^{(1)}$. This term includes the first approximation of the contribution to Φ_1 of couplings in the nonadiabatic region [cf. Eq. (39e)]. Although this term is second order in size, it is dipole in character. That is to say that it comes from a part of the wave function associated with the function $P_1(\cos\theta_{12})$ in the original expansion, Eq. (9).

In order to evaluate the integrals associated with the Eqs. (41) to (43), it is of course necessary to know the functions $\Phi_0^{(0)}$, $\Phi_1^{(0)}$, $\Phi_2^{(0)}$, and $\Phi_1^{(1)}$. The zeroth-order function $\Phi_0^{(0)}$ is effectively evaluated by methods of Section III. In obtaining this solution, we have used the separable nature of Eq. (39a). Reference to the same set of equations shows that none of the others are separable, thus the method we have used is inapplicable to them. We shall outline how, without solving the other equations outside of the adiabatic region, one can nevertheless get reasonable estimates of the integrals in question. The key idea here is the derivation of certain "sum rules," that is, the integral relationships between $\Phi_0^{(0)}$ and the higher $\Phi_l^{(j)}$. The point is that one can append a reasonable cutoff factor depending on an initially undetermined parameter to $\Phi_l^{(\mathrm{adiab})}$ of Eq. (15). The parameter is determined by satisfying the "sum rule." The (approximate) $\Phi_l^{(j)}$ can then be used to evaluate the desired multipole integrals of Eqs. (41) to (43).

The derivation of the sum rules follows closely the derivation of the basic relation Eq. (23). In this case, multiplying Eq. (39a) by $\Phi_1^{(0)}$ and Eq. (39b) by $\Phi_0^{(0)}$, subtracting, integrating over all $r_2 \leq r_1$, and using Green's theorem, one arrives at the dipole sum rule.

$$\int_0^\infty \int_0^{r_1} \Phi_0^{(0)} \left[-2\left(\frac{1}{r_1^2}+\frac{1}{r_2^2}\right)\right] \Phi_1^{(0)}\, dr_1\, dr_2 = \frac{2}{\sqrt{3}} \int_0^\infty \int_0^{r_1} (\Phi_0^{(0)})^2 (r^2/r_1^2)\, dr_2\, dr_1. \tag{44}$$

Similarly using Eqs. (39a) and (39d), multiplying on the left by $\Phi_2^{(0)}$ and $\Phi_0^{(0)}$, one obtains a quadrupole sum rule:

$$\int_0^\infty \int_0^{r_1} \Phi_0^{(0)} \left[-6\left(\frac{1}{r_1^2}+\frac{1}{r_2^2}\right)\right] \Phi_2^{(0)}\, dr_1\, dr_2 = \frac{2}{(5)^{1/2}} \int_0^\infty \int_0^{r_1} (\Phi_0^{(0)})^2 (r_2^2/r_1^3)\, dr_1\, dr_2. \tag{45}$$

Now for the triplet case consider the functions

$$\tilde{\Phi}_l^{(0)} = \frac{-2}{(2l+1)^{1/2}} \frac{\sin(kr_1+\delta_0)}{r_1^{l+1}} e^{-r_2} \left(\frac{r_2^{l+2}}{l+1}+\frac{r_2^{l+1}}{l}\right) \quad \{1-\exp[-D_l(r_1-r_2)]\}. \tag{46}$$

Independent of the value of the constant $D_l > 0$, the functions $\Phi_l^{(i)}(r_1 = r_2) = 0$ as has (tacitly) been assumed for the triplet solution of the appropriate one of the equations (39). [The tilde emphasizes that the functions defined by Eq. (46) are not identical with the solutions of Eq. (39)]. Let us additionally note that the cutoff factor does not change any of the other assumed boundary conditions of $\Phi_l^{(0)}$. The D_l are now determined by satisfying Eqs. (44) and (45) with $\Phi_l^{(0)}$ replaced by $\tilde{\Phi}_l^{(0)}$. The functions $\tilde{\Phi}_l^{(0)}$ are only meant to simulate $\Phi_l^{(0)}$ in the nonadiabatic region; that they can reasonably be expected to do so follows from the fact that they too satisfy the nonseparable boundary condition along $r_1 = r_2$, and hence are nonseparable functions. In Fig. 5 we have plotted the results of a dipole sum-rule calculation as explained in the

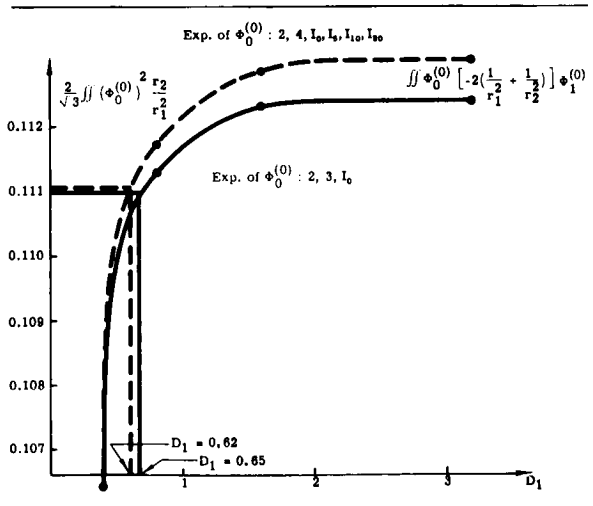


FIGURE 5.—Triplet dipole sum rule for $k = 0.1$ and two different expansions of $\phi_0^{(0)}$. The values of $\Delta\delta_0$ using the two expansions of $\phi_0^{(0)}$ and the corresponding values of D_l are 0.0303 and 0.0304, respectively.

caption. The differences coming from different (approximate) expansions of $\Phi_0^{(0)}$ are small compared to the intrinsic uncertainty of the sum-rule technique itself. There can be little question that the technique does take care of a large part of the nonadiabatic as well as the adiabatic contribution.

In the singlet case it would also be necessary to append a nonseparable function to $\Phi_l^{(\text{adiab})}$ in order that the singlet $\tilde{\Phi}_l^{(0)}$ obey the boundary condition $(\partial/\partial n \Phi_l^{(0)})_{r_1=r_2}=0$. However one can readily convince oneself that to derive such a truly one-parameter cutoff function is practically as difficult as solving the $\Phi_l^{(0)}$ equations. We have therefore contented ourselves with the Ansatz, in the singlet case

$$\tilde{\Phi}_l^{(0)}=\frac{-2}{(2l+1)^{1/2}}\frac{\sin(kr_1+\delta_0)}{(r_1{}^{l+1}+D_l)}e^{-r_2}\left(\frac{r_2{}^{l+2}}{l+1}+\frac{r_2{}^{l+1}}{l}\right). \tag{47}$$

The motivation of the cutoff factor is that $\tilde{\Phi}_l^{(0)}\to r_1{}^{l+1}$ in the limit $r_2=r_1\to 0$. This is the expected behavior of $\Phi_l^{(0)}$ if one neglects the coupling terms on the rhs of the equations. Figure 6 shows a typical quadrupole sum-rule graph. The value of D_2 was 8.65 practically independent of what "good" (see Section III for the definition of good) expansion of $\Phi_0^{(0)}$ was used.

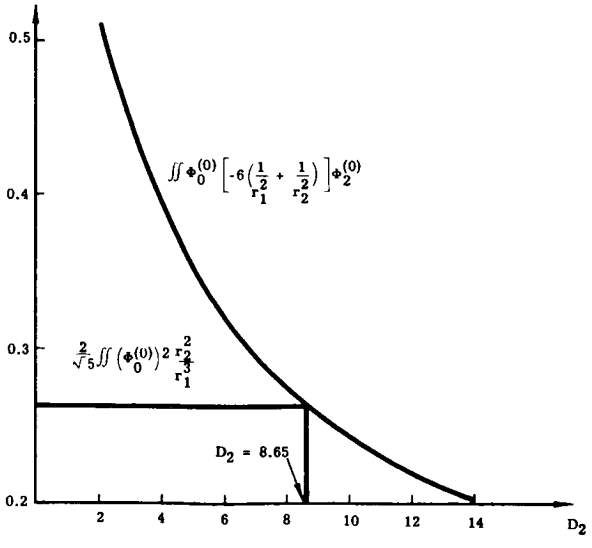


FIGURE 6.—Singlet quadrupole sum rule for $k=0.5$. Different expansions of $\phi_0^{(0)}$ give virtually indistinguishable curves on this scale. The value of $\Delta^2\delta_0^{(2)}$ is 0.0232.

The agreement, in fact, would tend to make one believe that the third figure of $\Delta^2\delta_0^{(2)}$ is significant. We do not believe this to be the case. To two figures, however, the answer is expected to be reliably independent of the cutoff. (That is, the result of using $\tilde{\Phi}_2^{(0)}$ instead of $\Phi_2^{(0)}$ is expected not to change the second significant figure.) Note that the cutoff that we have used leaves $\tilde{\Phi}_l^{(0)}$ separable; this is not as unreasonable as it appears, for the repulsion induced by the interaction is not as absolute as that induced by the exclusion principle. The latter operates in the previously discussed triplet case. Nevertheless because the cutoff there is more unambiguous and truly nonseparable, the third figure may have some significance. In both cases the cutoffs do not introduce any bending of the nodal lines in the r_1–r_2 plane. Since the amount of bending is expected to increase as the energy increases, the argument concerning significance must again be restricted to the lower energies.

In order to apply this kind of technique to the evaluation of $\Delta^2\delta_0^{(1)}$, one must know first the asymptotic form of $\Phi_1^{(1)}$. This latter may be found trivially in the adiabatic region from the realization that

$$\Phi_1^{(0)}+\Phi_1^{(1)}\underset{r_1\to\infty}{\longrightarrow}-\frac{2}{\sqrt{3}}\frac{\sin(kr_1+\delta_0+\Delta\delta_0)}{r_1{}^2}e^{-r_2}(\tfrac{1}{2}r_2{}^3+r_2{}^2)$$

This immediately yields

$$\Phi_1^{(1)(\text{adiab})}\underset{r_1\to\infty}{=}-\frac{2}{\sqrt{3}}(\Delta\delta_0)\frac{\cos(kr_1+\delta_0)}{r_1{}^2}e^{-r_2}(\tfrac{1}{2}r_2{}^3+r_2{}^2). \tag{48}$$

The functions $\tilde{\Phi}_1^{(1)}$ were constructed from the rhs of Eq. (48) by appending the same type of cutoff factors as in the previous two cases. One can derive sum rules for the determination of the parameters; however, these depend on not only the functions $\Phi_1^{(0)}$, $\Phi_0^{(0)}$, and $\Phi_2^{(0)}$ but on $\Phi_0^{(1)}$, $\Phi_1^{(1)}$ as well. It might be possible to derive additional sum rules for the latter such that one might find as many equations as unknown parameters, but the accumulation of cutoff factors quickly arrives at a point of no return. Therefore we chose the parameter as the mean between D_1 and D_2 from the previous two calculations and associated a much larger proportional error with it than with our calculation of $\Delta^2\delta_0^{(2)}$ or $\Delta\delta_0$. The uncertainty associated with $\Delta\delta_0$, however, because

it is generally an order of magnitude larger than the second-order corrections is expected to be larger than the error associated with the second-order terms.

The numerical integration must obviously be stopped at a finite point $r_1 = R$ and cannot be carried to infinity as formally required. In this connection we come again to difficulties of strictly numerical procedure which we now explicitly show must be augmented by an analytical device. The difficulty has to do with the fact mentioned in Section II that for the lowest energies the contribution from large distances has a nonvanishing effect on the phase shifts. Specifically there are long-range induced polarization forces which affect both the form and value of the effective range parameters. The changes in form have been dealt with elsewhere (Spruch and Rosenberg, 1960). We shall here treat the quantitative effect on the scattering length.

Our treatment will explicitly apply to the present case where both the zeroth order and exact phase shifts approach π from below; that is,

$$\lim_{k\to 0}\delta = \pi - ak$$

$$\lim_{k\to 0}\delta_0 = \pi - a_0 k.$$

On using this in Eq. (23), the basic relation for the scattering length becomes

$$a = a_0 + \frac{1}{k^2}\sum_{l=1}^{\infty}\frac{2}{(2l+1)^{1/2}} \times \int_0^{\infty}\int_0^{r_1} dr_1\, dr_2 \Phi_0^{(0)}(r_2^{l}/r_1^{l+1})\Phi_l.$$

The perturbation theoretic version of this formula has the sum of double integrals replaced by a double sum exactly as in Eq. (40). We shall concern ourselves only with first-order correction, which alone is significant to the accuracy we have attained. The machine evaluated an approximation of Δg_0 of Eq. (41)

$$\Delta\delta(R) = -k^{-1}\frac{2}{\sqrt{3}}\int_0^{R} dr_1 \int_0^{r_1} dr_2 \Phi_0^{(0)}(r_2/r_1^2)\Phi_1^{(0)}, \tag{49}$$

where, as we have stated the upper limit had to be replaced by a finite point $r_1 = R$. Consider the part of the integral beyond that point, that is,

$$\Delta\delta_0 - \Delta\delta(R) = -k^{-1}\frac{2}{\sqrt{3}}\int_R^{\infty} dr_1 \int_0^{r_1} dr_2 \Phi_0^{(0)}(r_2/r_1^2)\Phi_1^{(0)}.$$

Under the assumption that R is large enough for the various functions to be replaced by their asymptotic forms this reduces to

$$\Delta\delta_0 - \Delta\delta(R) = k^{-1}\frac{8}{3}\int_R^{\infty}\frac{\sin^2(kr_1+\delta_0)}{r_1^4} \times \int_0^{r_1} dr_2 e^{-2r_2}(\tfrac{1}{2}r_2^5 + r_2^4).$$

It is also assumed that R is large enough for the upper limit of the r_2 integral to be extendable to infinity without loss of accuracy. This gives

$$\Delta\delta_0 - \Delta\delta(R) = \frac{9}{2k}\int_R^{\infty}\frac{\sin^2(kr_1+\delta_0)}{r_1^4}\,dr_1. \tag{50}$$

We are concerned with this expression in the limit of zero energy. In order to find that limit, we again use the fact that

$$\lim k\to 0, \qquad \delta_0 \to \pi - a_0 k.$$

This can be inserted in the foregoing to give

$$\lim_{k\to 0} k^{-1}[\Delta\delta_0 - \Delta\delta(R)] = \frac{9}{2}\int_R^{\infty}\frac{(a_0 - r_1)^2}{r_1^4}\,dr_1 = \tfrac{9}{2}[R^{-1} - (a_0/R^2)] + 0(R^3)^{-1}$$

If we call $a(R)$ the scattering length associated with the calculation including all terms (even beyond the dipole) integrated to a point R, then

$$a(R) = \lim_{k\to 0}\frac{\sin\delta(R)}{k} = a_0 + \lim_{k\to 0}\frac{\Delta\delta(R)}{k} - \lim_{k\to 0}\frac{\Delta^2\delta^{(1)}(R)}{k} + \cdots,$$

then the foregoing leads to the formula for the true phase shift $a = a(\infty)$:

$$a = a(R) - \alpha[R^{-1} - (a_0/R^2)] + \cdots, \tag{51}$$

where $\alpha = \frac{9}{2}$ is the polarizability of atomic hydrogen. This formula is one of the most interesting

analytic results of the present method. Originally our calculation was carried out to $R=20$ and the associated $a(20)$ was found to be equal to 1.94, a quantity *above* the lower bound of Rosenberg *et al.* (1960). The result presented no small paradox in view of the inherent rigor of our method. When account was taken of the long-range polarization in the manner just described (Temkin 1961a), it led to a value $a=1.74$ which showed that the Rosenberg et al., bound of 1.91 was much further above the true scattering length than was previously thought to be the case.

At higher energies the integral on the rhs of Eq. (50) can be carried out numerically to a sufficiently large value of r_1 for convergence to take place. Analytically one can make an expansion in terms of (generally) transcendental functions of k. The first term is proportional to k and leads to an alteration of the usual form of effective range expansion to (Spruch et al. 1961).

$$k\cot\delta=-a^{-1}+(\pi\alpha/3a^2)k+O(k^2\ln k).$$

The linear term in k is completely absent from the effective range expansion for short-range potentials. As far as the expansion of the integral in Eq. (50) is concerned, we found that the first two terms contributed only 90% to the integral as numerically determined for k as low as 0.01. This indicates that the series for $k\cot\delta$ is slowly convergent, if indeed it is convergent at all and not asymptotic.

We now come to the numerical aspects of the above calculations. It is clear that ultimately one must integrate a set of double integrals. This involves dividing the triangular region of Fig. 1 into a suitably small lattice and evaluating the integrands at each of the lattice points. Specifically a square lattice of mesh size h was introduced. Each column along r_1 was integrated using the trapezoidal rule and the resulting row of numbers was again integrated with the trapezoidal rule. For integrals that vanish along all boundaries, this amounts to summing the integrand at all interior mesh points and multiplying by h^2. (In spite of the crudeness of this integrating scheme, doubling the mesh from $h=0.05$ to $h=0.1$ generally affected only the fourth significant figure.)

Now consider an integration to a maximum $r_1=20$ with a mesh size $h=0.05$. This amounts to approximately 80000 points at which an integrand must be evaluated. The most complicated function, which enters all integrands, is $\Phi_0^{(0)}$, the expansion of which we repeat here.

$$\Phi_0^{(0)}=\sin(kr_1+\delta_0)R_{1s}(r_2)+\sum_j C_{nj}\exp(-\kappa_{nj}r_1)R_{nj}(r_2)+\sum_j C_{pj}\exp(-\kappa_{pj}r_1)\mu_{pj}(r_2).$$

The quantities C_{nj}, C_{pj}, and δ_0 are known from the zeroth-order calculation. Let us estimate the time necessary to calculate $\Phi_0^{(0)}$ for an arbitrary mesh point (r_1, r_2). Take on the average an expansion of $\Phi_0^{(0)}$ consisting of four terms, each term consisting of three factors. Confining ourselves to a discrete state $R_{ns}(r)$, so that

$$R_{ns}(r)=e^{-r/n}\sum_{t=1}^{n}c_{nt}r^t$$

with an average $n=3$, one can determine that it takes 14 multiplications, 4 division, 2 additions, 1 square root, and 1 exponential to compute $R_{3s}(r)$. The approximate IBM 7090 floating arithmetic times are

Operation	Time in microseconds = (μ sec)
Addition	6.4
Multiplication	11
Division	13
Exponential	340
Square Root	212

This yields 758 μ sec to compute $R_{3s}(r)$. One can similarly estimate that it takes 588 μ sec to compute $\exp(-\kappa_3 r_1)$, so that it takes approximately 1.3 msec to compute one term of $\Phi_0^{(0)}$. If there are on the average four terms in $\Phi_0^{(0)}$, then it requires 5 msec to compute them: add an additional millisecond to compute the remaining factors in a typical integrand and one arrives at 6 msec per point. To compute a double integral of 80,000 points, therefore, it would take about 8.4 min. We estimate that in the process of debugging, trying different cutoff factors, searching for equality in sum rules, and computing the higher order terms themselves, that we computed of the order of two thousand integrals. Had we used the method we have just described to compute the

```
745    DI = 0.
750    DO 830 I = 1, NS
755    RI = FLOATF(I)*H
765    SUUM = 0.
770    DO 800 J = 1, I
775    RJ = FLOATF(J)*H
778    SUM = F(1, I, 1)*F(1, J, 2)
780    DO 790 K = 2, NN11
790    SUM = F(K, I, 1)*F(K, J, 2)*C(K - 1) + SUM
798    SUMM = SUM*F(NN11 + 1, I, 1)*F(NN11 + 1, J, 2)*RJ
800    SUUM = SUMM + SUUM
803    SUUM = -SUMM*.5 + SUUM
830    DI = SUUM + DI
850    DBINT = (DI-.5*SUUM)*H** 2 * 4./(3.*CAY)
```

FIGURE 7.—Facsimile of program to compute $\Delta\delta_0^{(R)}$ of Eq. (49). NS is the number of mesh points along r_1 (same as along r_2). NN11 is the number of terms in $\Phi_0^{(0)}$. H is the mesh size. $C(K)$, $K=1,\cdots$, NN11−1 are the coefficients which multiple the 2d through NN11st terms of $\Phi_0^{(0)}$. $F(K, I, M)$ are the tabulated values ($I=1, \cdots$, NS of the $k=1, \cdots$, NN11 functions which enter $\Phi_0^{(0)}$. For $M=1$ they refer to the functions of r_1; for $M=2$ they refer to the functions of r_2. $F(\mathrm{MM11}+1, I, M)$ are the functions $\sin(kr_1+\delta_2)/(r_1^2+D)r_1^2$ and $e^{-r2}(-r_2^3+r_2^2)$ for $M=1$, 2, respectively. CAY$=k$, and DBINT$=\Delta\delta_0^1(R)$ where $R=H\times$NS.

integrands, we would have required over two hundred forty hours of machine time. This is a lot of machine time. It is possible with some clever programming that one could cut this time down by a factor of two or even three. However, without any ingenuity, but with a little common sense, one can easily save a factor of ten. The idea is to construct tables of the various functions at all the mesh points in one dimension and store them. In the example above this would mean computing eight functions of four hundred points each. This takes, at most, a couple of seconds of computer time. To construct each term of $\Phi_0^{(0)}$ then requires two multiplications of the appropriate terms in the various tables or eight multiplications and three additions (that is, 108 μsec) in all to construct a point of $\Phi_0^{(0)}$. The remaining factors in any integrand when similarly handled will bring the total time per point to $\leq$500 μsec which is less than a tenth of the previous value. Of course, this latter method requires more storage cells than the former, but this was readily available in our 32,000 magnetic core storage.

We have gone into this long discussion to emphasize an obvious but important point: *in multiple integrals (or any multidimensional problem) it is essential to minimize the operating time per point.* As a corollary, the example also illustrates when storage is available it should be completely subservient to time as a quantity to be minimized.

In Fig. 7 we have given an excerpt for the program evaluating the singlet $\Delta\delta(R)$ of Eq. (49). The point there is that although it can take a lot of computer time, the programming only takes a few lines.

The remaining problem of numerical interest is the determination of the continuum Coulomb functions, $u_p(r)$, which are also necessary to evaluate $\Phi_0^{(0)}$ at any given lattice point. The construction of tables of Coulomb wave functions has occupied both physicists and numerical analysts for many years. The difficulty lies in the fact that the various numerical techniques for evaluating these functions do not yield sufficient accuracy except for a small region of the relevant parameters. Thus, in order to construct a table of sufficient accuracy covering an adequate range of the relevant parameters, one has had to use a vertiable jigsaw puzzle of expansions, integral representation, recursion relations, etc.

In our problem no great accuracy is required, three or four places being entirely sufficient for the purposes of double integration. Therefore the main consideration was the utilization of one technique applicable to all cases. After some thought it was decided that direct numerical integration of the differential equation for $u_p(r)$, Eq. (25a) was the most convenient way of achieving this. The differential equations contain no second derivative which makes them admirably suited to the numerical technique of Hartree (1958). The method utilizes the fact that the second difference of the solution $u(r)$ of such an equation is related to successively higher even differences of the second derivative of the same function,

$$\delta^2 u_0 = (\delta x)^2[u_0'' + \tfrac{1}{12}\delta^2 u_0'' - \tfrac{1}{24}\delta^4 u_0'' + \cdots], \qquad (52)$$

where for any function $f(r)$ and mesh size h

$$f_n \equiv f(r+mh)$$

and

$$\delta^2 f_n \equiv f_{n+1} - 2f_n + f_{n+1}.$$

TABLE VI.—*Exact Versus Numerically Integrated Values of* $u_p(r)$

	$u_0(r)=(\frac{1}{2}r)^{\frac{1}{2}}J_1(8r)^{\frac{1}{2}}$			$u_1(r)$	
	Exact[a]	Numerical integration $h=0.05$[c]	Numerical integration $h=0.01$	Exact[b]	Numerical integration $h=0.05$
1	0.282986	0.282980	0.282975	0.2077802	0.207800
2	−0.0660433	−0.0660439	−0.066058	−0.2352408	−0.2352410
3	−0.385239	−0.385239	−0.385245	−0.3018270	−0.3018265
4	−0.465153	−0.465150	−0.465127	0.0781332	0.0781344
5	−0.318969	−0.318964	−0.318905	0.362463	0.362463
6	−0.045650	−0.045648	−0.0455614	—	—
7	+0.245234	0.245236	0.245325	—	—
8	+0.4692726	0.4692744	0.469328	—	—

[a] Calculated from Watson (1953).
[b] Calculated from Abramowitz (1952).
[c] h is the mesh size of the numerical integration.

If one knows u_0, u_{-1}, u_{-2}, and u_0'', u_{-1}'', and u_{-2}'' then the idea is to take a guess at δ^2u_0'' (say $\delta^2u_0''=\delta^2u_{-1}''$, which is known) and calculate δ^2u_0 and hence $u_1=u(r+h)$ from the above equation. (One retains only two terms in Eq. (52).) One can use the differential equation itself to compute u_1'' and hence a new δ^2u_0''. If the difference between the old and new δ^2u_0'' is small enough, then the u_1 can be considered as a satisfactory approximation of $u(r+h)$, and, one can go on to compute $u(r+2h)$. If the difference is not small enough, one can use the new δ^2u_0'' to compute a new u_1 and iterate the process until δ^2u_0'' does converge, before going on to the next point. Fortunately this technique has been programmed and is available in the SHARE library[3] of subroutines (SDA 141, LAS 887). As programmed, the integration routine does not advance until the difference between successive δ^2u_0'' is less than 10^{-6}.

We found in actually using this subroutine that loops tended to develop in which presumably successive values of δ^2u_0'' were oscillating between two values which were further apart than 10^{-6}. We therefore altered the program so that a given point was never iterated more than ten times. The pragmatic justification of this modification was that no matter how peculiar the behavior of $u_p(r)$ at an isolated point r, this would not affect the general behavior of $u_p(r)$ to the low accuracy we required, particularly for the purposes of quadrature.

For small values of p, where one could readily make a comparison table VI shows a comparison of our results with essentially tabulated results. For $p=0$ we used in column of "exact" entries the formula

$$u_0(r)=(r/2)^{1/2}J_1(8r)^{1/2}$$

and the tabulated values in Watson (1958) with interpolation done via a Taylor series. The superiority of the results for $h=0.05$ as compared with $h=0.01$ is presumably associated with round-off errors. For our best results the agreement is so good that the technique of numerical integration looks promising as a means of computing tables of Coulomb wave functions. It should be realized, however, that the usefulness of tables is being constantly altered with the emerging availability of high-speed computing machines to a means of checking machine calculations of specific functions.

RESULTS AND DISCUSSION

In Tables VII and VIII we present results of the zeroth-order calculation. The results represent only a small sample of those actually obtained.

TABLE VII.—*Triplet Results for $\Phi_0^{(0)}$ and δ_0*

k				Exchange
	Exp[a]	2, 3, I_0	2, 4, I_0	2, 4, I_0, I_5, I_{10}, I_{20}	3.118
0.01	δ_0 (rad)	3.11820	3.11823	3.11821	
	I_T	-0.165×10^{-7}	-0.149×10^{-7}	-0.424×10^{-8}	
		2, 3, I_0	2, 4, I_0	2, 4, I_0, I_5, I_{10}, I_{20}	3.024
0.05		3.0247	3.0255	3.0257	
		0.160×10^{-6}	0.168×10^{-6}	0.244×10^{-5}	
		2, 3, I_0	2, 4, I_0	2, 4, I_0, I_5, I_{10}, I_{20}	2.907
0.1		2.9084	2.9097	2.9102	
		0.283×10^{-6}	0.465×10^{-5}	0.699×10^{-5}	
		2, 3, I_0	2, 4, I_0	2, 4	2.679
0.2		2.6806	2.6807	2.6809	
		0.153×10^{-5}	0.173×10^{-5}	0.283×10^{-5}	
		2, 3, I_0	2, 4, I_0	2, 5, I_0, I_5, I_{10}, I_{20}	2.461
0.3		2.4634	2.4629	2.4632	
		0.722×10^{-5}	0.433×10^{-5}	0.549×10^{-5}	
		2, 3, I_0	2, 4, I_0, I_5, I_{10}, I_{20}	2, 4, I_0, I_3, I_5, I_7, I_{10}, I_{20}	2.257
0.4		2.2582	2.2588	2.2589	
		0.113×10^{-5}	0.201×10^{-5}	0.169×10^{-5}	
		2, 4, I_0	2, 4, I_0, I_5, I_{10}, I_{20}	2, 4, I_0, I_3, I_5, I_7, I_{10}, I_{20}	2.070
0.5		2.0715	2.0716	2.0721	
		0.477×10^{-5}	0.398×10^{-5}	0.365×10^{-5}	
		2, 4, I_0	2, 3, 4, I_0, I_1, I_2, I_3, I_7	2, 3, 4, I_0, I_1, I_2, I_3	1.679
0.75		1.6914	1.6830	1.6830	
		0.355×10^{-3}	0.497×10^{-4}	0.474×10^{-4}	
		2, 3, 4, I_0, I_1, I_2, I_3	2, 3, 4, I_0, I_1, I_2, I_3, I_7	2, 4, I_0, I_3, I_5, I_7, I_{10}, I_{20}	1.614
0.8		1.61664	1.61665	1.6220	
		-0.117×10^{-4}	-0.123×10^{-4}	0.18×10^{-2}	

[a] The first line of each entry refers to the terms used in the expansion of $\Phi_0^{(0)}$. The second line is δ_0 in radians. The third line is the diagonal sum I_T.

Table IX contains the zero-order phase shifts together with the higher order corrections. The uncertainty in the last figure (given by the number in parentheses) was estimated from the agreement of the various δ_0 with each other from Tables VII and VIII.

The important property of the numbers in the succeeding columns is their obvious convergence. First we note $\Delta\delta_0$ is small compared to δ_0. (From the point of view of the cross section; however, the correction should be compared with $\pi-\delta_0$ for $\delta_0>\pi/2$). Secondly, we note that the second-order correction, $\Delta^2\delta_0^{(1)}+\Delta^2\delta_0^{(2)}$, is just about an order of magnitude less than the first-order correction. This constitutes an *a posteriori* demonstration of the convergence of the multipole series and of our perturbation theory. On the basis of this convergence we feel that we can meaningfully ascribe an error of our δ from the exact result. This is also given as the number in parentheses in the appropriate coulumns. The convergence appears to be such that one could confidently

TABLE VIII.—*Singlet Results for* $\Phi_0^{(0)}$ *and* δ_0

k						Exchange
	Exp	3, 4	4, I_0	3, I_0, I_3	3, 4, I_3, I_4	3.0606
0.01	δ_0 (rad)	3.0638	3.0637	3.0636	3.;643	
	I_S	-0.373×10^{-8}	-0.186×10^{-7}	-0.198×10^{-7}	0.242×10^{-7}	
		4, I_0	4, I_0, I_4	3, I_0	3, I_0, I_3	
0.05		2.7595	2.7596	2.7591	2.7593	2.746
		-0.238×10^{-6}	0.263×10^{-6}	0.157×10^{-6}	0.149×10^{-6}	
		5, I_0, I_5	4, I_0, I_2, I_3, I_4	4, I_0, I_3	4, I_0, I_4, I_8, I_{12}, I_{16}	
0.1		2.4207	2.4174	2.4182	2.4177	2.396
		0.774×10^{-6}	0.783×10^{-6}	0.918×10^{-6}	0.934×10^{-6}	
		2, 3, I_1, I_2, I_3, I_4	2, 5, I_1, I_2, I_3, I_4	2, I_0, I_1, I_2, I_3, I_4	3, 4, I_1, I_2, I_3	
0.2		1.8949	1.8968	1.8947	1.8960	1.870
		0.777×10^{-5}	0.142×10^{-4}	0.193×10^{-4}	0.253×10^{-4}	
		2, 5, I_1, I_2, I_3, I_4	2, I_0, I_1, I_2, I_3, I_4, I_5		2, 3, I_0, I_6	
0.3		1.5350	1.5321		1.5245	1.508
		0.183×10^{-4}	0.291×10^{-4}		0.295×10^{-4}	
		2, 4, I_1, I_2, I_3	2, 5, I_1, I_2, I_3, I_4	2, I_0, I_1, I_2, I_3, I_4	2, 3, I_2, I_3	
0.4		1.2694	1.2691	1.2685	1.2697	1.239
		0.489×10^{-6}	0.150×10^{-5}	0.693×10^{-5}	0.134×10^{-4}	
		3, 4, I_1, I_2	3, 5, I_1, I_2, I_3, I_4	2, I_0, I_1, I_2, I_3, I_4, I_5	2, I_0, I_2, I_4, I_6, I_8	
0.5		1.0667	1.0652	1.0647	0.0656	1.031
		0.926×10^{-5}	0.353×10^{-4}	0.375×10^{-4}	0.459×10^{-4}	
		2, 3, 4, $I_{0.75}$, $I_{1.5}$, $I_{2.25}$, I_3, $I_{3.75}$, $I_{4.5}$	2, 3, I_0, $I_{0.5}$, I_1, $I_{1.5}$, I_2	2, 3, 4, $I_{0.5}$, I_1, $I_{1.5}$, I_2, $I_{2.5}$	2, I_0, $I_{0.5}$, I_1, $I_{1.5}$, I_2, $I_{2.5}$	
0.75		0.7556	0.7564	0.7535	0.7566	0.694
		0.517×10^{-5}	0.117×10^{-4}	0.109×10^{-4}	0.131×10^{-4}	
		2, 4, 5, $I_{0.75}$, $I_{1.5}$, $I_{2.25}$	2, 3, I_0, $I_{0.5}$ I_1 $I_{1.5}$ I_2, $I_{2.5}$	2, 3, 4, $I_{0.75}$, $I_{1.5}$ $I_{2.25}$ I_3 $I_{3.75}$ $I_{4.5}$	2, 3, 5, $I_{0.5}$ I_1 $I_{1.5}$, I_2, $I_{2.5}$	
0.8		0.7289	0.7268	0.7261	0.7273	0.651
		0.170×10^{-4}	0.175×10^{-4}	0.212×10^{-4}	0.228×10^{-4}	

TABLE IX.—*Resume of Calculation for Singlet and Triplet Phase Shifts*

	Triplet					Singlet				
k	δ_0	$\Delta\delta_0$	$\Delta^2\delta_0^{(1)}$	$\Delta^2\delta_0^{(2)}$	δ	δ_0	$\Delta\delta_0$	$\Delta^2\delta_0^{(1)}$	$\Delta^2\delta_0^{(2)}$	δ
0[a]	2.338 (3)				1.76 (3)	7.8 (1)				5.6 (4)
0.01	3.11821(2)	0.00504	0.000295	0.000153	3.1237(4)	3.0640(5)	0.026	−0.008	0.004	3.086(4)
0.05	3.025 (1)	0.0193	0.0013	0.0008	3.046 (5)	2.759 (1)	0.117	−0.035	0.019	2.86 (2)
0.1	2.909 (1)	0.0303	0.0021	0.0014	2.942 (10)	2.420 (3)	0.187	−0.045	0.030	2.59 (3)
0.2	2.6810 (5)	0.0379	0.0021	0.0023	2.723 (10)	1.895 (2)	0.215	−0.030	0.034	2.11 (5)
0.3	2.4630 (5)	0.0392	0.0017	0.0026	2.516 (10)	1.535 (10)	0.189	−0.016	0.030	1.74 (6)
0.4	2.259 (1)	0.0379	0.0013	0.0028	2.301 (10)	1.269 (1)	0.165	−0.009	0.026	1.45 (6)
0.5	2.072 (1)	0.0363	0.0011	0.0027	2.112 (10)	1.066 (2)	0.148	−0.007	0.023	1.23 (6)
0.75	1.683 (2)	0.0328	0.0007	0.0024	1.719 (10)	0.756 (2)	0.131	−0.0025	0.021	0.91 (6)
0.8	1.617 (2)	0.0282	0.0005	0.0018	1.647 (10)	0.728 (2)	0.126	−0.002	0.020	0.87 (6)

[a] The $k=0$ entries are scattering lengths.

expect that the higher order terms will not contribute more than 5 to the fifth figure. The errors we have indicated are greater than this because they include the uncertainties in our calculations of δ_0, $\Delta\delta_0$, and the second-order corrections. The largest of these in magnitude is that associated with $\Delta\delta_0$, as was stated in the last section. In the triplet case we have compared our result with some preliminary results based on numerical integration of the partial differential equations. The deviation near $k=0$ appears to be less than 5 percent. As k increases, the particular form of cutoff we have used is less reliable, as we have also mentioned in the last section. We have, therefore, taken the triplet error to range from 5 percent to 25 percent increasing linearly with k. In the singlet case we have taken the error in $\Delta\delta_0$ to range from 10 percent to 40 percent. The final estimate of the error, we feel, is very liberal; however, the possibility of doing better with the encouraging results of our recently developed program in numerical integration persuades us not to push our present results farther than we have. In addition, and most important, the phase shifts that we have already computed should be quite adequate for experimental purposes.

There have been many calculations of electron-hydrogen scattering too numerous to discuss in their entirety here. Of the earlier ones, the most important is that of Morse and Allis (1933). As it applies to S-wave scattering their method (which nowadays is known as the exchange-approximation) is equivalent to a variational approximation of our zeroth-order problem, since it does not include any dependence on the angle θ_{12}. It turns out that the phase shifts so derived are in excellent agreement with δ_0 (cf. last column of tables VII and VIII).

Very recently Rosenberg et al. (1960) have computed singlet and triplet scattering lengths of 6.22 and 1.97 (in units of Bohr radii), respectively. These calculations are important because they represent rigorous upper bounds on the exact scattering lengths. This is the first time that anything completely rigorous about the electron-hydrogen problem was ever knowlingly calculated. In fact, aside from the same authors' positron-hydrogen calculation (Spruch and Rosenberg, 1960) which again is based on their important proof that certain expressions do give upper bounds for the scattering length, this is the first time that anything rigorous has ever been said concerning any many-body scattering problem. Their calculation, however, did not represent any kind of systematic search for a minimum.

Such a search has been carried out by Charles Schwartz (1961), contemporaneously with our

TABLE X.—*Comparison of Schwartz' (1961) Phase Shifts with Present Results*

k	Triplet		Singlet	
	Schwartz	Non-adiabatic	Schwartz	Non-adiabatic
0.0	1.7686(2)	1.76 (3)	5.965 (3)	5.6 (4)
0.1	2.9388(4)	2.942(10)	2.553 (1)	2.59(3)
0.2	2.7171(5)	2.723(10)	2.0673(9)	2.11(5)
0.3	2.4996(8)	2.516(10)	1.6964(5)	1.74(6)
0.4	2.2938(4)	2.301(10)	1.4146(4)	1.45(6)
0.5	2.1046(4)	2.112(10)	1.202 (1)	1.23(6)
0.8	1.643 (3)	1.647(10)	0.886 (1)	0.87(6)

own work, (cf. the article by C. Schwartz in this volume). In table X we display his results against our own. His search consisted of using a Hylleraas-type wave function in Kohn's variational principle and constantly increasing the number of terms. His estimate of the error is based on the convergence of his minimized results as a function of the number of terms. His calculations include up to 50 terms.

First, it can clearly be seen that within our (larger) estimated error the results agree. However, there is a tendency for our phase shifts to be slightly larger than his. In an effort to understand the discrepancy between our results, Dr. Schwartz has done a calculation with only $P_0(\cos\theta_{12})$ (functions involving r_1 and r_2 only), and another calculation with $P_0(\cos\theta_{12})$ and $P_1(\cos\theta_{12})$ terms only, both for $k=0.4$ in the triplet case. Comparing P_0 and P_0+P_1 results with his final answers, he concludes that dipole makes a contribution 0.0340 to the phase shift. All the remaining multipoles contribute 0.0015. The dipole contributions in our calculation is given by $\Delta\delta_0+\Delta^2\delta_0{}^{(1)}$ and can be taken from the appropriate row of table IX; one gets 0.0392; the quadrupole alone adds an additional 0.0028 to our phase shift. Now as we have stated, the sum rule technic itself could possibly account for the difference in the dipole contributions; however, it is difficult to imagine that our $\Delta^2\delta_0{}^{(2)}$ could be off by a factor of 2. If the error lies in our calculation, it could mean that the contributions from successive multipoles do not converge as rapidly as one thinks or that the contributions to a given multipole coming from our perturbation theory, do not converge well. We think that the first of these possibilities is unlikely, but that the second is not as unlikely. Therefore unless there is a more fundamental theory it is possible that our convergence is misleading. In the absence of this, one must be guided by physical intuition. We think that when our methods of numerical integration are perfected, so that we can eliminate the uncertainty in the contributions of the included $\Phi_l{}^{(j)}$ to the phase shifts, that we can obtain an additional figure of accuracy.

There is likewise a possibility that the apparent convergence of Schwartz' results are misleading. With a view to suggesting a more definite method, it is perhaps appropriate that we should mention certain reservations concerning a Hylleraas-type wave function in connection with a scattering problem. First it should be noted that to the Hylleraas form that is used in the bound-state problem, one must append a nonvanishing sinusoidal term whose coefficient determines the scattering one is trying to calculate. Aside from this term all terms vanish exponentially in all asymptotic regions of configuration space. However, correctly we know that in the adiabatic region there are slowly vanishing sinusoidal components of the true wave function (that is, Φ_l, $l>0$). The direct contribution of this region of configuration space to the phase shifts above say, 0.5 ev can be shown to be very small. Nevertheless one can imagine that the mere requirement of this distorted boundary condition may impose a form of the wave function in the region of interaction which may not be well approximated by a certain form of trial function. (This argument is not unrelated to the difficulties of numerical integration mentioned in Section II.)

More specifically we have in mind the following: a ground-state wave function can be considered a surface in the space of the independent variables. In a certain sense it is the "smoothest" allowable surface. This accounts for the fact that one can get reasonably accurate energies with a comparatively few parameters. On the other hand a

wave function corresponding to a scattering problem constitutes a relatively "bumpy" surface. If, therefore, one attempts to construct a replica of this surface using the same type of functions, one may require tremendously more parameters to do it. In other words, the convergence may be slow indeed. It is precisely the residue of this "bumpiness" which ebbs into both the nonvanishing and slowly vanishing sinusoidal components of the true wave function.

We therefore think that from the point of view of presently attainable accuracy, a variational calculation with, say, a Hylleraas-type wave function for the interior and variational counterparts of the $\Phi_l^{(\mathrm{adiab})}$ to represent the adiabatic boundary condition offers the most unassailable technique. The adiabatic terms should include at least up to quadrupole parts, and they should not be confined to zero energy. In addition they should contain variational parameters of their own to take care of any peculiar interior behavior of a scattering as opposed to a bound-state wave function.

At zero energy the situation is different in that the higher multipoles especially the dipole contribute directly and significantly ($\sim 10\%$) to the scattering length (Temkin 1961a). This was discussed in connection with the derivation of Eq. (51). The result was subsequently found by Schwartz (1961), who, when he added the counterpart of $\Phi_1^{(\mathrm{adiab})}$ to his wave function, obtained a corresponding lowering of his scattering lengths.* For this reason and for the reason that his zero-energy results are absolute upper bounds, there can be little question that his scattering lengths are the most accurate yet calculated. At higher energies, the comparative accuracy is more uncertain.

The question of accuracy concerning phase shifts should be tempered by the realization that scattering parameters can only be poorly experimentally determined as compared to bound-state energies. From the point of the cross-section measurements, differences between the results in table X are completely negligible. It is more important, in fact, to have accurate values of the higher partial wave phase shifts. The partial differential equations whose solutions yield these are, however, much more complicated than the S-wave equation to which we have here confined ourselves.

---

* Refer to the abstracts by A. Temkin and C. Schwartz in the Proceedings of the "2nd International Conference on the Physics of Electronic and Atomic Collisions," (1961). It will be noted that our final triplet scattering length in table IX is slightly different (although within the stated error) from the above reference and Temkin (1961a). The difference comes from the fact that Eq. (44) is not strictly valid when integrating over a finite region. Since we did, in fact, integrate over a finite region a correction had to be inserted. This turns out to be significant only in the above noted case.

Appendix I*

Evaluation of the matrix element

$$(\mathrm{ME})_{I_{p1},I_{p2}}=\int_0^\infty e^{-\lambda r}u_{p_1}(r)\,u_{p_2}(r)\,dr.$$

The $u_p(r)$ are the continuum Coulomb wave functions

$$\begin{aligned}u_p(r)&=re^{-ipr}F(1+ip^{-1};\,2;\,2ipr)\\&=ia^{-1}(e^{-i\beta r}/2i)\,(\partial/\partial p)F(ia;\,1;\,2ipr),\end{aligned}$$

where F is a confluent hypergeometric function. In the second line we have used the derivative property of F and have let $\beta=p=1/a$ in order to emphasize that the derivative operates only on the last argument. We now use the contour integral representation of F (Nordsieck, 1954)

$$F(ia;\,1;\,z)=(2\pi i)^{-1}\oint dt\,t^{ia-1}(t-1)^{-ia}e^{zt},$$

where the contour goes in the counterclockwise direction around the branch cut connecting the two branch points at $t=0$ and $t=1$ in the complex t plane. We therefore can use this in the integral and perform the r integration to get

$$(\mathrm{ME})_{I_{p1},I_{p2}}=\frac{1}{16\pi^2(ia_1)\,(ia_2)}\,\frac{\partial^2}{\partial p_1\partial p_2}\oint\oint dt_1\,dt_2 t_1^{ia_1-1}\frac{(t_2-1)^{-ia_2}}{\lambda+i\beta_1+i\beta_2-2ip_1t_1-2ip_2t_2}$$

The integrand now has a pole in the t_2 plane at

$$t_2=(2ip_2)^{-1}(\lambda+i\beta_1+i\beta_2-2ip_1t_1).$$

One can deform the contour around the branch line in the t_2 plane to enclose this pole in the counterclockwise direction. This is a well-known trick used, for instance, to evaluate the photoeffect matrix element (cf. Bethe and Salpeter, 1957, p. 304), and in this case allows us to perform the t_2 integration. The residue yields

$$(\mathrm{ME})_{I_{p1},I_{p2}}=\frac{i}{8\pi(ia_1)\,(ia_2)}\,\frac{\partial^2}{\partial p_1\partial p_2}\oint\frac{dt_1}{t_1}\,\frac{(1-t_1^{-1})^{-ia_1}\{1-[2ip_2/(\lambda+i\beta_1+i\beta_2-2ip_1t_1)]\}^{-ia_2}}{\lambda+i\beta_1+i\beta_2-2ip_1t_1}$$

The additional singularities in the remaining integral have an imaginary part $-2\lambda/p_1<0$, so that the contour around $t_2=0,\,1$ can be made not to include them. We now make the transformation $t_1=1/t$; the transformed contour in the t plane will not include the branch points at $t_2=0,\,1$ but can include the other two singularities. We thus get

$$(\mathrm{ME})_{I_{p1},I_{p2}}=\frac{i}{8\pi(ia_1)\,(ia_2)}\,\frac{\partial^2}{\partial p_1\partial p_2}\oint dt\frac{(1-t)^{-ia_1}[t(\lambda+i\beta_1+i\beta_2)-2ip_1]^{-1+ia_2}}{[(\lambda+i\beta_1+i\beta_2-2ip_2)t-2ip_1]^{ia_2}}$$

* We are greatly indebted to Dr. L. Maximon for the contents of both Appendix I and Appendix II.

Letting.

$$\alpha=\lambda+i\beta_1+i\beta_2,$$

$$\delta=\beta=2ip_1, \qquad \text{(do not confuse with } \beta_{1,2}=p_1, p_2)$$

$$\gamma=\lambda+i(\beta_1+\beta_2-2p_2),$$

$$\tau=(\gamma t-\delta)/(\gamma-\delta),$$

we can write the integral in the above expression

$$\oint=\left(\frac{\gamma-\delta}{\gamma}\right)\left(\frac{\gamma-\delta}{\gamma}\right)^{-ia^1}\left(\frac{\alpha(\gamma-\delta)}{\gamma}\right)^{-1+ia^1}(\gamma-\delta)^{-ia_2}\oint d\tau(1-\tau)^{-ia_1}(\tau-x)^{-1+ia_2}\tau^{-ia_2},$$

where

$$x=[(\beta\gamma-\alpha\delta)/\alpha(\gamma-\delta)].$$

The τ integral is the integral representation of $2\pi iF(1-ia_2, ia_1; 1; x)$, so that we have

$$(\mathrm{ME})_{Ip_1,Ip_2}=\frac{i}{8\pi(ia_1)(ia_2)}\frac{\partial^2}{\partial p_1\partial p_2}\left\{(2\pi i)\frac{\gamma^{i(a_1-a_2)}\alpha^{-1+ia_2}}{(\gamma-\delta)^{ia_1}}F(1-ia_2, ia_1; 1; x)\right\}.$$

To perform the differentiation is a messy, but straightforward, task. One has to use the differential equation satisfied by the hypergeometric function to get the resultant expression in a convenient form. One gets finally

$$(\mathrm{ME})_{Ip_1,Ip_2}=\frac{2\lambda(\alpha/\gamma)^{ia_2}[\alpha/(\gamma-\delta)]^{ia_1}}{\gamma\gamma^*\alpha(\gamma-\delta)}\left[V-\frac{2y}{\gamma^*}W\right], \tag{53}$$

where

$$y=1-x$$

and

$$V=F(-ia_2, ia_1; 1; x)$$

$$W=F(1-ia_2, 1+ia_1; 2; x).$$

The substitution of the various quantities in terms of λ, p_1 and p_2 yields the expression given in Eq. (36).

Appendix II

Proof of the reality and symmetry of $(\mathrm{ME})_{Ip_1,Ip_2}$.

We start with the matrix element in the form (53) of Appendix I. Taking the complex conjugate we have

$$(\mathrm{ME})^*_{Ip_1,Ip_2}=\frac{2\lambda}{\alpha\alpha^*\gamma\gamma^*}\left(\frac{\alpha^*}{\gamma^*}\right)^{-ia_2}\left(\frac{\gamma^*}{\alpha}\right)^{-ia_1}\left[V^*-\frac{2y^*}{\gamma}W^*\right].$$

Now one can show from the properties of the hypergeometric functions that

$$W^*=y^{i(a_1-a_2)}W$$

and

$$V^*=y^{i(a_1-a_2)}[V+i(a_1-a_2)xW]$$

One therefore gets, using $y=\gamma\gamma^*/(\alpha\alpha^*)$,

$$(\mathrm{ME})^*_{Ip_1,Ip_2}=\frac{2\lambda}{\alpha\alpha^*\gamma\gamma^*}\left(\frac{\alpha^*}{\gamma^*}\right)^{-ia_2}\left(\frac{\gamma^*}{\alpha}\right)^{-ia_1}\left(\frac{\gamma\gamma^*}{\alpha\alpha^*}\right)^{i(a_1-a_2)}\left[V+i(a_1-a_2)xW-\frac{2\gamma^*}{\alpha\alpha^*}W\right],$$

which can be reduced further to

$$(\mathrm{ME})^*_{Ip_1,Ip_2}=\frac{2\lambda}{\alpha\alpha^*\gamma\gamma^*}\left(\frac{\gamma}{\alpha^*}\right)^{ia_1}\left(\frac{\alpha}{\gamma}\right)^{ia_2}\left[V+\left\{i(a_1-a_2)x-\frac{2\gamma^*}{\alpha\alpha^*}\right\}W\right].$$

The quantity in parenthesis that is multiplied by W can be shown to be equal to $-2\gamma/\alpha\alpha^*$, such that

$$(\mathrm{ME})^*_{Ip_1,Ip_2}=\frac{2\lambda}{\alpha\alpha^*\gamma\gamma^*}\left(\frac{\gamma}{\alpha^*}\right)^{ia_1}\left(\frac{\alpha}{\gamma}\right)^{ia_2}\left[V-\frac{2\gamma}{\alpha\alpha^*}W\right]$$

$$=(\mathrm{ME})_{Ip_1,Ip_2}. \qquad \text{Q.E.D.}$$

The proof of symmetry (S), that is, invariance under the operation $1\rightleftarrows 2$, follows from the above proven invariance under complex conjugation (C) and the manifest invariance under the combined operation of S and C. [This latter invariance may visually be ascertained from the matrix element in the form of Eq. (36)]. Symbolically the latter property may be written

$$\mathrm{CS}(\mathrm{ME})=\mathrm{SC}(\mathrm{ME})=(\mathrm{ME}).$$

But we have just proved $\mathrm{C}(\mathrm{ME})=\mathrm{ME}$. Thus it follows that

$$\mathrm{S}(\mathrm{ME})=\mathrm{ME}.$$

REFERENCES

ABRAMOWITZ, M. "Tables of Coulomb Wave Functions," Volume I. United States Government Printing Office, Washington, D. C., 1952.

BETHE, H. A., and SALPETER, E. E. "Quantum Mechanics of One- and Two-Electron Atoms," Springer, Berlin, 1957.

HARTREE, D. R. "Numerical Analysis." Clarendon Press (Oxford), London, 1958.

LUKE, P., MEYEROTT, R., and CLENDENIN, W. *Phys. Rev.* **85,** 401, 1952.

MORSE, P. M., and ALLIS, W. P. *Phys. Rev.* **44,** 269, 1933.

MORSE, P. M., and FESHBACH, H. "Methods of Theoretical Physics." McGraw-Hill, New York, 1953.

NORDSIECK, A. *Phys. Rev.* **93,** 785, 1954.

ROSENBERG, L., SPRUCH, L., and O'MALLEY, T. F. *Phys. Rev.* **119,** 164, 1960.

SCHWARTZ, C. *Phys. Rev.* **124,** 1468, 1961.

SPRUCH, L., and ROSENBERG, L. *Phys. Rev.* **117,** 143, 1960.

SPRUCH, L., ROSENBERG, L., and O'MALLEY, T. F. *Phys. Rev. Letters* **5,** 375, 1961.

TEMKIN, A. *Phys. Rev. Letters* **4,** 566, 1960.

TEMKIN, A. *Phys. Rev. Letters* **6,** 354, 1961a.

TEMKIN, A. *J. Math. Phys.* **2,** 336, 1961b.

TEMKIN, A. *Phys. Rev.* **126,** 130, 1962.

WATSON, G. N. "A Treatise on the Theory of Bessel Functions." Cambridge Univ. Press, London and New York.

NONADIABATIC THEORY OF ELECTRON-HYDROGEN SCATTERING, PART II

A. TEMKIN AND E. SULLIVAN

Goddard Space Flight Center

The triplet S-wave electron-atomic hydrogen elastic scattering phase shifts are recalculated by a previously introduced nonadiabatic theory. The previous calculation has been improved in a number of respects, the most important of which is the use of a noniterative technique for numerically solving the partial differential equations. (This technique is expected to be useful for a large class of linear second order elliptic partial differential equations.) Phase shifts are computed to better than four significant figures. The results are quite close to the variational results of Schwartz but on the whole somewhat larger. The deviations are considered significant, and the various approaches are discussed: Specifically our triplet scattering length (in Bohr radii) is $a_t = 1.7683$.

INTRODUCTION

The extension of the relative partial wave treatment to the (electron-hydrogen) scattering problem was introduced[1] to allow for the calculation of phase shifts of sufficient accuracy for experimental purposes and to allow for meaningful comparison by approximate theories. The completion of the original program[2] has apparently met the purposes for which it was intended.[3] Reference 2 will be referred to here as Part I, and equations in it which are referred to will be prefixed by a I. This includes any minor variations in notation.

With regard to the original calculation, it was clear from the first that the devices introduced to elicit information about the higher corrections—which, precisely speaking, involved the solutions of two-dimensional partial differential equations—limited the accuracy to significantly less than that to which the method was intrinsically capable. In addition, a variational calculation[4] has appeared in which the estimated accuracy was much higher than in reference 2.

The variational calculation employed (in Kohn's variational principle) a Hylleraas-type wave function with an increasingly large number N of parameters. The estimate of the error was based on the device, first exploited by Pekeris,[5] of observing the results as a function of N. However, for at least two reasons, the variational calculations pertaining to scattering are not compelling to the accuracy claimed. First, the variational results at nonzero energies show, as a function of the nonlinear parameter, a kind of wild behavior which has required a very intuitive method of interpretation.[6] (For $k>0$ we do not have the cushion of a guaranteed lower bound on the phase shifts.) More important, however, is the fact that a Hylleraas wave function does not naturally describe the complete wave function corresponding to a scattering problem. In fact, it has been conclusively demonstrated[7] that at zero energy the long-range adiabatic tail is essential for highly quantitative purposes. Although the variational calculation finally did include such a term at zero energy, it did not include it for nonzero energies; it is still very much in question to what extent this term enters at small but finite energies.

For these reasons it has seemed necessary to carry out our intention of numerical integration for the higher order correction. The calculation has been restricted to the triplet case as discussed in the next section. In the third section of this

paper we discuss the method of numerically integrating the partial differential equations. Finally in the fourth section of this paper we present results and discussion.

REVIEW OF THE NONADIABATIC THEORY

It will be recalled that the nonadiabatic theory[2] starts with a decomposition of the S-wave function

$$\Psi\left(r_1\, r_2\, \theta_{12}\right) = \frac{1}{r_1\, r_2} \sum_{l=0}^{\infty} \sqrt{2l+1}\; \Phi_l\left(r_1 r_2\right)\, P_l\left(\cos\theta_{12}\right), \tag{I-3}$$

from which by substitution into the Schrodinger equation an infinite set of coupled two-dimensional partial differential equations (equations I-4) results. We define a zeroth order problem by neglecting the coupling terms of the $l=0$ equation; thus

$$\left(\Delta_{12} + E + \frac{2}{r_2}\right) \Phi_0^{(0)}\left(r_1\, r_2\right) = 0, \tag{I-11}$$

where

$$\Delta_{12} \equiv \frac{\partial^2}{\partial r_1^2} + \frac{\partial^2}{\partial r_2^2}.$$

The zeroth order wave function $\Phi_0^{(0)}$ is required to have the asymptotic boundary condition corresponding to a scattered wave:

$$\lim_{r_1 \to \infty} \Phi_0^{(0)}\left(r_1\, r_2\right) = \sin\left(kr_1 + \delta_0\right)\, R_{1s}\left(r_2\right), \tag{I-12}$$

where δ_0 is the zeroth-order phase shift, and δ_0 can be interpreted as the phase shift of a rudimentary type of three-body problem, to which, it has been shown in reference 2, many previous approximations were unknowingly addressed.

The basic relation of the nonadiabatic theory is given by

$$\sin\left(\delta - \delta_0\right) = -\frac{1}{k} \sum_{l=1}^{\infty} \frac{2}{\sqrt{2l+1}} \int_0^{\infty} dr_1 \int_0^{r_1} \Phi_0^{(0)}\, \frac{r_2^{\,l}}{r_1^{\,l+1}}\, \Phi_l\, dr_2, \tag{I-13}$$

where δ is the exact S-wave phase shift. The integration domain is confined to the region $r_1 \geq r_2$. The convergence of the terms on the right-hand side was established by noting that the significant contribution to each integral comes from two regions. One is the adiabatic region, $r_1 \gg r_2$ and r_2 small. We shall have much to say about this contribution in connection with the shortcomings of the Hylleraas wave function variational approach; however, for the purposes of the present nonadiabatic theory these contributions can readily be accounted for, and can be shown to diminish rapidly as a function of l. The other region which must be considered is for intermediate values r_1 and r_2; this gives the essential contribution to the deviation of δ and δ_0. A powerful (but nonrigorous) argument for the convergence of this contribution has been given in reference 2. Briefly it was noted there that each Φ_l equation was a centrifugal barrier term $-l(l+1)(r_1^{-2}+r_2^{-2})$ which successively diminishes the amplitude of Φ_l in that region. This argument is the analog of the well known argument that the contributions of successive partial waves go down for low impacting energies in the complete description of the scattering process. It should be emphasized, however, that there is not a one-to-one correspondence in this analogy, for in the equations of partial waves there are no coupled terms, whereas in equation I-4 the relative partial wave equations are coupled. Thus, whereas we can assert rigorously that only partial S-waves contribute to the zero-energy cross section, it is not true that only the s-relative angular momentum state con-

tributes to the S-wave phase shift at zero energy. What is here being asserted is that the contributions from higher relative angular momentum states diminish in a usefully convergent manner.

There is an additional circumstance, which was not really emphasized in reference 2, which renders the argument of convergence particularly cogent in the triplet case. In that case we have the boundary conditions

$$\Phi_l\ (r_1 = r_2) = 0, \quad l = 0, 1, \cdots$$

$$\Phi_0^{(0)}\ (r_1 = r_2) = 0 \cdot \tag{I-6}$$

From inspection of equation I-13 we can see, by virtue of the $r_2{}^l/r_1{}^{l+1}$ factor, that the region $r_2 \lesssim r_1$ would tend to become increasingly important for the higher relative partial waves. However, because of the boundary condition, the contribution from this region must in fact be small; therefore the main contribution to each integral must come from r_2 significantly less than r_1. But in that region $r_2{}^l/r_1{}^{l+1}$ certainly diminishes rapidly as a function of l. Thus, we have every reason for believing the convergence will be exceedingly rapid in the triplet case. The same arguments should also render our perturbation theory

$$\Phi_l = \sum_{j=0}^{\infty} \lambda^{j+l/2}\, \Phi_l^{(j)}$$

particularly effective.

These reasons plus the calculated values which are presented below give a strong expectation that when the terms through quadratic order are included in the calculation, more than four-place accuracy will be obtained in the phase shifts. This accuracy is required if the results are to be compared meaningfully with Schwartz's results.

The original calculation[2] has been improved in two main respects. Firstly we have generalized the zeroth order technique of solution to double precision arithmetic on the IBM 7090 computer. In the single precision program we were plagued with vanishing determinants which very rapidly used up all the eight significant figures that the machine could store. In the double precision program, the 16 significant figures were sufficient to yield zeroth order phase shifts from better than four to almost eight significant figures.

To illustrate the gain in accuracy of our zeroth order results by making the program double precision, table 1 presents a typical set of results for $k=0.2$. The middle two columns refer to quantities labeled det

TABLE I.—*Zeroth Order Double Precision Results for* $k=0.2$

Expansion	Determinant	Diagonal Sum	δ_0
2	0.154×10^{-1}	0.358×10^{-2}	2.71098
2,3	0.315×10^{-6}	0.141×10^{-4}	2.679565
2,3,4	0.541×10^{-14}	0.231×10^{-6}	2.6794215
2,3,4,5	0.418×10^{-25}	0.137×10^{-8}	2.6794197
2,3,4,5,I_0	0.246×10^{-36}	0.553×10^{-7}	2.67962
2,3,I_0	0.621×10^{-12}	0.823×10^{-7}	2.6794200
2,3,I_0,I_1	0.329×10^{-16}	0.142×10^{-7}	2.6794192
2,4,I_0	0.123×10^{-12}	0.773×10^{-8}	2.6794191
2,4,I_0,I_5,I_{10},I_{20}	0.638×10^{-26}	0.742×10^{-8}	2.6794191

and I_T in reference 2. From the first of these we can obtain an indication of the number of significant figures lost in the evaluation of the determinant det. For example, in the row corresponding to the expansion 2, 3, 4, and 5, the main diagonal of det, that is

$$\prod_{i=1}^{4}(ME)_{ii},$$

is approximately 0.6×10^{-9}, if we figure 0.5×10^{-2} as the average value of a diagonal matrix element. Comparing this value with 0.42×10^{-20}, the actual value of the determinant, we see that approximately 16 significant figures have been lost;* this is the maximum number of significant figures that the double precision arithmetic affords, and the conclusion is reinforced by referring to the next row, 2, 3, 4, 5, I_0. Here almost 25 significant figures have been lost, and the corresponding diagonal sum increases! If the elements of the calculation have sufficient (infinite) accuracy, it is clear that the addition of a term to an expansion can only decrease the diagonal sum. Thus, the δ_0 for that expansion is very unreliable. In each of the groups of rows separated by dotted lines each subsequent row augments the previous one by an additional term. It should be noted that except for the above case the diagonal sums diminish. This should be compared with table 3 of reference 8 in which a similar set of results based on a single precision program was presented. There an increase in the diagonal sum occurred in third row! In the present case, we could with some legitimacy claim $\delta_0=2.6794194(3)$; however the value $\delta_0=2.67942$ is quite adequate for our purposes.

The second and main advancement we have achieved in the way of computing the complete S-wave phase shifts is the numerical integration of the higher order equations. These are elliptic equations, and it is well known that the usual way of numerically solving an elliptic equation is by some sort of relaxation of iteration technique. The novel feature of the technique we have used is that it is not iterative. Because it has worked where a relaxation technique has utterly failed to converge, and because it is applicable to a whole class of linear second-order equations, we have given some detail in the next section to the numerical solution.

SOLUTION OF THE PARTIAL DIFFERENTIAL EQUATIONS

The higher order effects in our expansion are given by the formula

$$\sin(\delta-\delta_0) = -\frac{1}{k}\sum_{\nu=1}^{\infty}\lambda^{\nu}\sum_{\substack{m+\mu=\nu\\ m\geq 1,\ \mu\geq 0}}\frac{2}{\sqrt{2m+1}}\int_0^{\infty}\int_0^{r_1}\Phi_0^{(0)}\,\frac{r_2^{m}}{r_1^{m+1}}\,\Phi_m^{(\mu)}\,dr_1\,dr_2 \qquad \text{(I-43)}$$

(For the purposes of the discussion in the fourth section of this paper we emphasize that this λ-expansion is a modification of the ordinary l-expansion in relative partial waves, which is expected to hasten the convergence in the triplet case. It also has additional advantages discussed in reference 2.)

The partial differential equation for $\Phi_0^{(0)}$ has been repeated in the last section. We repeat here the

* The formula for tan2δ involves, among other things, the value of the determinant. It does not follow, however, that there is a one-to-one correspondence between the number of significant figures in the determinant and in tan2δ. We have inferred the number of significant figures by the uniformity of the results for δ as shown in table 1. The details are discussed further in reference 8.

remaining equations:

$$\left[\Delta_{12} - \frac{2}{r_1^2 + r_2^2} + \frac{2}{r_2} + E\right]\Phi_1^{(0)} = \frac{2}{\sqrt{3}}\frac{r_2}{r_1^2}\Phi_0^{(0)}, \tag{I-39}$$

$$\left[\Delta_{12} + \frac{2}{r_2} + E\right]\Phi_0^{(1)} = \frac{2}{\sqrt{3}}\frac{r_2}{r_1^2}\Phi_1^{(0)}, \tag{I-40}$$

$$\left[\Delta_{12} - \frac{6}{r_1^2 + r_2^2} + \frac{2}{r_2} + E\right]\Phi_2^{(0)} = \frac{2}{\sqrt{5}}\frac{r_2^2}{r_1^3}\Phi_0^{(0)}, \tag{I-41}$$

$$\left[\Delta_{12} - \frac{2}{r_1^2 + r_2^2} + \frac{2}{r_2} + E\right]\Phi_1^{(1)} - \frac{4}{5}\frac{r_2^2}{r_1^3}\Phi_1^{(0)} = \frac{2}{\sqrt{3}}\frac{r_2}{r_1^2}\Phi_0^{(1)} + \frac{4}{\sqrt{15}}\frac{r_2}{r_1^2}\Phi_2^{(0)}. \tag{I-42}$$

In principle these equations are to be integrated over the infinite region $0 \leq r_2 \leq r_1 < \infty$. In practice, of course, we can only integrate up to a finite point $r_1 = R$. If R is suitably large, we can perform the quadratures in equation I-43 from R to ∞ by replacing the functions by their adiabatic forms:

$$\Phi_0^{(0)} \approx \sin(kr_1 + \delta_0)\, R_{1s}(r_2), \tag{I-17}$$

$$\Phi_l^{(0)} \approx -\frac{2\sin(kr_1 + \delta_0)}{r_1^{l+1}\sqrt{2l+1}}(e^{-r_2})\left(\frac{r_2^{l+2}}{l+1} + \frac{r_2^{l+1}}{l}\right), \qquad l = 1, 2 \tag{I-47}$$

$$\Phi_1^{(1)} \approx -\frac{2}{\sqrt{3}}(\Delta\delta_0)\frac{\cos(kr_1 + \delta_0)}{r_1^2}(e^{-r_2})\left(\frac{r_2^3}{2} + r_2^2\right). \tag{I-48}$$

(These functions are exact solutions in the limit $r_1 \to \infty$.)

The δ_0 being known, these equations with $r_1 = R$ then define the boundary conditions on the numerical solution in the region $0 \leq r_2 \leq r_1 \leq R$. It must be emphasized that as long as R is finite there is an approximation involved in these equations as boundary conditions. For these forms require not only $R \to \infty$, but also that $r_2 \ll R$. However, for numerical purposes we must know the boundary conditions for all values of $r_2 \leq R$: the most natural thing to do is simply to use the above equations for all values of r_2. If R is large enough, we can be sure that the error thus incurred is very small, because in the region $r_2 \lesssim R$ both the above functions and the true functions will be negligibly small.

The important question concerning R is, how large is large enough? The time required for numerical integration effectively limited us to an $R = 10$. It was our original intention to use the numerically integrated $\Phi_0^{(0)}$ in the quadratures (equation I-43) together with the remaining numerically integrated functions. Our suspicion of the accuracy of the numerically integrated function was aroused when at $k = 0.4$ we found that

$$\int_0^R \int_0^{r_1} \frac{r_2}{r_1^2}\left(\Phi_0^{(0)}\right)^2 dr_1\, dr_2$$

was half the value that we had gotten using the analytic expansion of $\Phi_0^{(0)}$ that we had as a by-product of the calculation of δ_0. We were therefore led to examine the deviation of our original $\Phi_0^{(0)}$

$$\Phi_0^{(0)} = \sin(kr_1 + \delta_0)\, R_{1s}(r_2) + \left(\sum_n + \int dp\right) C_n e^{-\kappa_n r_1} R_{ns}(r_2) \tag{I-19}$$

TABLE 2.—*Comparison of the Numerically Integrated Boundary Values at* $r_1=10$ $(k=0.4)$ *with the Expansion of* $\Phi_0^{(0)}$ *in Equation* I-19.

r_2	$\Phi_0^{(0)}(10,r_2)$	$\sin(10k+\delta_0)\,R_{1s}(r_2)$
1	-0.01913	-0.01853
2	-0.01363	-0.01363
3	-0.00685	-0.00752
4	-0.00259	-0.00369
5	-0.00042	-0.00170
6	0.000486	-0.000749
7	0.000722	-0.000321
8	0.000612	-0.000135
9	0.000334	-0.0000056
10	-0.000003	0[a]

[a]The expression on the left is not 0 at this point; however, for the purposes of numerical integration this point, being equal to r_1 at r_2, it was automatically taken as 0 by our program.

from the boundary values imposed on the numerical solution by equation I-17. The results are given in table 2.

The important thing to notice is that the values obtained from the analytic $\Phi_0^{(0)}$ along the line $r_1=10$ change sign at $r_2\approx 5$ whereas those defined by equation I-17, being proportional to $R_{1s}=2r_2\exp(-r_2)$, do not. We concluded that this change of sign which was not taken care of in the boundary condition of equation I-17 was responsible for the inaccuracy of the numerically integrated $\Phi_0^{(0)}$.

The double integral in which we observed the discrepancy is of course part of the dipole "sum rule"

$$\int_0^{\infty}\int_0^{r_1}\Phi_0^{(0)}\left[-2\left(\frac{1}{r_1^2}+\frac{1}{r_2^2}\right)\right]\Phi_1^{(0)}\,dr_1\,dr_2 \;=\; \frac{2}{\sqrt{3}}\int_0^{\infty}\int_0^{r_1}\left(\Phi_0^{(0)}\right)^2\frac{r_2}{r_1^2}\,dr_1\,dr_2\,. \tag{I-51}$$

In the original calculation[2] we used this relation to find the nonadiabatic effects of $\Phi_1^{(0)}$. In the present case, equation I-51 together with its counterpart for $\Phi_2^{(0)}$, plus additional sum rules that we can derive, serve as a check of the numerical integration. Now in the above noted $k=0.4$ case it was observed that the $\Phi_1^{(0)}$ found from equation I-39 using the numerically integrated (incorrect) $\Phi_0^{(0)}$ gave approximate equality in the sum rule. At work here was undoubtedly the phenomenon that the (incorrect) $\Phi_0^{(0)}$ is an important enough inhomogeneous term in the differential equation to influence $\Phi_1^{(0)}$ to be incorrect in just such a way as to give equality in equation I-51. This experience prevented any complacency on our part that the satisfaction of sum rules of the numerical function was a foolproof guarantee that the functions were correct.

The $k=0.4$ case is extreme in the sense that at no other energy have we observed the analytic $\Phi_0^{(0)}$ to change sign in such a prominent place along the boundary. (The change apparently stems from the fact that $\Phi_0^{(0)}$ has a node very close to $r_1=10$ at this energy.) Accordingly the differences between the integrals on the right-hand side of equation I-51 using the numerical versus the analytic $\Phi_0^{(0)}$ at other energies differed only in the second significant figure. However, since we require practically three significant figures in the integral

$$\Delta\delta_0 \;=\; -\frac{1}{k}\left(\int_0^{10}\int_0^{r_1}\Phi_0^{(0)}\,\frac{r_2}{r_1^2}\,\Phi_1^{(0)}\,dr_1\,dr_2+\int_{10}^{\infty}\int_0^{r_1}\Phi_0^{(0)}\,\frac{r_2}{r_1^2}\,\Phi_1^{(0)}\,dr_1\,dr_2\right) \tag{I-44}$$

we could not use the numerically integrated $\Phi_0^{(0)}$.

We therefore decided to use the $\Phi_0^{(0)}$ obtained from the best expansion (equation I-19) we had and numerically integrate for the remaining functions. Our unfortunate experience with the dipole sum rule

gave us some confidence that having a reliable $\Phi_0^{(0)}$ would make up for any deficiencies in the boundary conditions for the other functions. The sum rules, of course, still are a necessary condition to be satisfied; however, in view of the necessity of cutting off the numerical integrals at R, the sum rules did not afford a critical test of the accuracy. We found that we did have to go a mesh size $h=0.1$ to get reasonable results. The most natural further tests of accuracy—further halving the mesh size, integrating out to double the value of R etc.—were precluded by machine storage and particularly time considerations. As it was, it took 90 minutes to integrate a differential equation on the IBM 7090, so that at nine energies and five equations per energy, we had already used a vast amount of time. As we have said, we are quite confident that we have attained a requisite accuracy, but our present inability to carry out further checks provides the most serious weakness in the present calculation. We hope to carry through some of these checks as bigger and faster computers become available.

We now turn to the problem of numerically integrating the partial differential equations. In the most general case, we have the following boundary value problem: $\phi(x, y)$ satisfies equation 1 in a domain $\mathfrak{D}$

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + g(x,y)\,\phi = f(x,y) , \tag{1}$$

where the values of ϕ on the boundary are known. If we replace the second order partial derivatives in equation 1 with the second differences

$$\frac{\partial^2 \phi}{\partial x^2} \cong \frac{1}{h^2}\left[\phi(x+h,y) - 2\phi(x,y) + \phi(x-h,y)\right] ,$$

$$\frac{\partial^2 \phi}{\partial y^2} \cong \frac{1}{h^2}\left[\phi(x,y+h) - 2\phi(x,y) + \phi(x,y-h)\right] ,$$

then equation 1 can be written as

$$\phi(x+h,y) + \phi(x-h,y) + \phi(x,y+h) + \phi(x,y-h) + \left[h^2\, g(x,y) - 4\right]\phi(x,y) = h^2\, f(x,y) .$$

(The method we shall subsequently present is readily extendable to higher order difference formulas.) When this is done we have reduced the problem to a solution of the matrix equation

$$A\phi = k , \tag{2}$$

with A an $N2 \times N2$ non-singular, real matrix of the coefficients of ϕ at the $N2$ internal mesh points and k a known column vector of the boundary values.

The $N2$ being on the order of the square of the number of mesh points along a boundary N, it appears that a direct inversion of equation 2 would be almost impossible. For that reason an iterative solution is usually attempted in which A is written

$$A = I - B ,$$

where I is the unit matrix.

The iteration procedure is defined by

$$\phi^{(n+1)} = B\phi^{(n)} + k . \tag{3}$$

The superscripts refer to the iteration number. The $\phi^{(0)}$ is an initial arbitrary guess of the solution. If we define λ_i, $i=1, n$ to be the eigenvalues of matrix B, then

$$\lambda_{max} = \max_{1 \le i \le n} |\lambda_i| .$$

A necessary and sufficient condition for the convergence of system of equation 3 is that $\lambda_{max} < 1$.[9]

The iterative techniques used in the solution of matrix equations are applicable to a wide class of boundary value problems, but it is sometimes very difficult to determine whether λ_{max} is less than one. If λ_{max} is less than but close to one, the number of iterations required for a satisfactory solution becomes quite large.

The following is an exact method of solving the matrix problem defined by equation 1. In essence we are exploiting the fact that although A is of gigantic dimension, it is of a special form, and, most of its its elements are zero.

The matrix A of equation 2 is of the form:

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & 0 & \cdots & & & & & & & & & \\ A_{2,1} & A_{2,2} & A_{2,3} & 0 & \cdots & & & & & & & & \\ 0 & A_{3,2} & A_{3,3} & A_{3,4} & & 0 & \cdots & & & & & & \\ \cdots & & & & & & & & & & & & \\ \cdots & 0 & A_{N-2,N-3} & A_{N-2,N-2} & A_{N-2,N-1} & 0 & & & & & & & \\ \cdots & 0 & A_{N-1,N-2} & A_{N-1,N-1} & A_{N-1,N} & & & & & & & & \\ \cdots & 0 & A_{N,N-1} & A_{N,N} & & & & & & & & & \end{bmatrix}$$

where for the particular triangular boundary of the zeroth order problem $A_{1,1}$ is of dimension $N \times N$: $A_{1,2}$ of $N \times (N-1)$; $A_{2,1}$ of $(N-1) \times N$; $A_{2,2}$ of $(N-1) \times (N-1)$; ... $A_{N,N-1}$ of 1×2 and $A_{N,N}$ of 1×1.

We now factor matrix A into the product of two matrices

$$A = LU ,$$

which in terms of block matrices can be written:

$$\begin{bmatrix} A_{1,1} & A_{1,2} & 0 & \cdots & & & & & & & & & \\ A_{2,1} & A_{2,2} & A_{2,3} & 0 & \cdots & & & & & & & & \\ 0 & A_{3,2} & A_{3,3} & A_{3,4} & & 0 & \cdots & & & & & & \\ \cdots & & & & & & & & & & & & \\ \cdots & 0 & A_{N-2,N-3} & A_{N-2,N-2} & A_{N-2,N-1} & 0 & & & & & & & \\ \cdots & 0 & A_{N-1,N-2} & A_{N-1,N-1} & A_{N-1,N} & & & & & & & & \\ \cdots & 0 & A_{N,N-1} & A_{N,N} & & & & & & & & & \end{bmatrix} \begin{bmatrix} L_{1,1} & & & & & & & & & & & & \\ L_{2,1} & L_{2,2} & & & & & & & & & & & \\ 0 & L_{3,2} & L_{3,3} & & & & & O & & & & & \\ \cdots & & & & & & & & & & & & \\ & & & 0 & L_{N-2,N-3} & L_{N-2,N-2} & & & & & & & \\ & & & & 0 & L_{N-1,N-2} & L_{N-1,N-1} & & & & & & \\ & & & & & 0 & L_{N,N-1} & L_{N,N} & & & & & \end{bmatrix} \begin{bmatrix} I & U_{1,2} & 0 & \cdots & & & & & & & & & \\ & I & U_{2,3} & 0 & \cdots & & & & & & & & \\ & & I & U_{3,4} & 0 & \cdots & & & & & & & \\ & & & & & & & & & & & & \\ & & & & & I & U_{N-2,N-1} & 0 & & & & & \\ & & & & O & & I & U_{N-1,N} & & & & & \\ & & & & & & & I & & & & & \end{bmatrix}$$

(4)

Performing the multiplication of L by U as defined by equation 4 we can express $L_{i,j}$, $U_{i,j}$ $i,j=1, N$ in terms of the given matrix A.

$$\begin{aligned} L_{1,1} &= A_{1,1} \\ L_{i,i-1} &= A_{i,i-1} \qquad i = 2,N \\ U_{i-1,i} &= L^{-1}_{i-1,i-1} A_{i-1,i} \qquad i = 2,N \\ L_{i,i} &= A_{i,i} - L_{i,i-1} U_{i-1,i} \qquad i = 2,N. \end{aligned} \tag{5}$$

The factorization of A in the form of equation 4 is only possible if the matrices $L_{i,i}$, $i=1, N$ are nonsingular.[10]

With this factorization, we have from equation 2

$$A\phi = LU\phi = k . \tag{6}$$

We let $U\phi=\zeta$, and this gives from equation 6

$$L\zeta = k , \tag{7}$$

$$\zeta = L^{-1}k . \tag{8}$$

Writing equation 7 more explicitly, we have in terms of the block matrices of L

$$\begin{bmatrix} L_{1,1} & & & & & & \\ L_{2,1} & L_{2,2} & & & & & \\ 0 & L_{3,2} & L_{3,3} & & & & \\ \cdot & \cdot & \cdot & & & & \\ & & 0 & L_{N-2,N-3} & L_{N-2,N-2} & & \\ & & & 0 & L_{N-1,N-2} & L_{N-1,N-1} & \\ & & & & 0 & L_{N,N-1} & L_{N,N} \end{bmatrix} \begin{bmatrix} \bar{\zeta}_1 \\ \bar{\zeta}_2 \\ \bar{\zeta}_3 \\ \cdot \\ \cdot \\ \cdot \\ \bar{\zeta}_N \end{bmatrix} = \begin{bmatrix} \bar{k}_1 \\ \bar{k}_2 \\ \bar{k}_3 \\ \cdot \\ \cdot \\ \cdot \\ \bar{k}_N \end{bmatrix} \tag{9}$$

where $\bar{\zeta}_1$, $\bar{k}_1$ are vectors of N elements, $\bar{\zeta}_2$, $\bar{k}_2$ are vectors of $(N-1)$ elements, ..., $\bar{\zeta}_N$, $\bar{k}_N$ are vectors of 1 element.

From equation 9:

$$\bar{\zeta}_1 = L_{1,1}^{-1} \bar{k}_1 ,$$

$$\bar{\zeta}_i = L_{i,i}^{-1} \left(\bar{k}_i - L_{i,i-1} \bar{\zeta}_{i-1} \right) \quad i = 2,N . \tag{10}$$

We have

$$U\phi = \zeta , \tag{11}$$

and putting this in more explicit form, we have

$$\begin{bmatrix} I & U_{1,2} & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & I & U_{2,3} & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & I & U_{3,4} & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & & I & U_{N-2,N-1} & 0 & & \\ & & & & & & & I & U_{N-1,N} & & \\ & & & & & & & & I & & \end{bmatrix} \begin{bmatrix} \bar{\phi}_1 \\ \bar{\phi}_2 \\ \bar{\phi}_3 \\ \cdot \\ \cdot \\ \cdot \\ \bar{\phi}_N \end{bmatrix} = \begin{bmatrix} \bar{\zeta}_1 \\ \bar{\zeta}_2 \\ \bar{\zeta}_3 \\ \cdot \\ \cdot \\ \cdot \\ \bar{\zeta}_N \end{bmatrix} \tag{12}$$

From equation 12 we obtain

$$\bar{\phi}_N = \bar{\zeta}_N ,$$

$$\bar{\phi}_{i-1} = \bar{\zeta}_{i-1} - U_{i-1,i} \cdot \bar{\phi}_i \qquad i = 2, N . \tag{13}$$

An examination of equations 10 and 13 indicates that the only inversions necessary in the entire computation are the inversions of the matrices $L_{i,i}$ $i=1, N$. This is a result of our factorization which places the identity matrix on the block diagonal matrices of U. It is of the essence that the matrices $L_{i,1}$ are, in practice, of a much smaller dimension than the original matrix. For the particular boundary we are considering (see also, figure 1 of reference 2) the dimension of $L_{i,i}$ decreases from $L_{N,N}$ (an $N \times N$ matrix) to $L_{1,1}$ (dimension 1×1).

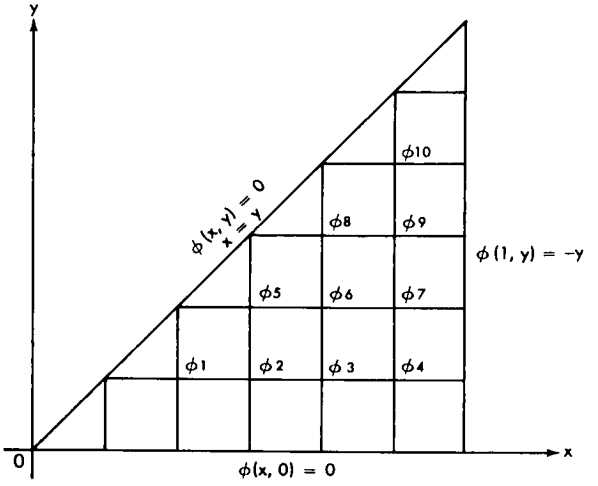


FIGURE 1.—The domain of integration ($\mathfrak{D}$) of the example given in equation 14.

We now give an example of the numerical method discussed above. Suppose we wish to solve the following boundary value problem in the domain $\mathfrak{D}$:

$$\frac{\partial^2}{\partial x^2}\phi(x,y) + \frac{\partial^2}{\partial y^2}\phi(x,y) +$$

$$\left(\frac{2}{y} - 6x - 1\right)\phi(x,y) = \frac{xe^{-x}}{36} . \tag{14}$$

We wish to determine the interior points $\phi_1, \phi_2, \phi_3, \ldots, \phi_{10}$ with the boundary values as specified in

figure 1. The problem is reduced to a solution of the following system of equations:

$$
\begin{aligned}
\tfrac{135}{36}\phi_1 + \phi_2 &= \tfrac{1}{3}e^{-1/3},\\
\phi_1 - \tfrac{136}{36}\phi_2 + \phi_3 + \phi_5 &= \tfrac{1}{2}e^{-1/2},\\
\phi_2 - \tfrac{137}{36}\phi_3 + \phi_4 + \phi_6 &= \tfrac{2}{3}e^{-2/3},\\
\phi_3 - \tfrac{138}{36}\phi_4 + \phi_7 &= \tfrac{1}{6} + \tfrac{5}{6}e^{-5/6},\\
\phi_2 - \tfrac{142}{36}\phi_5 + \phi_6 &= \tfrac{1}{2}e^{-1/2},\\
\phi_3 + \phi_5 - \tfrac{143}{36}\phi_6 + \phi_7 &= \tfrac{2}{3}e^{-2/3},\\
\phi_4 + \phi_6 - \tfrac{144}{36}\phi_7 &= \tfrac{1}{3} + \tfrac{5}{6}e^{-5/6},\\
\phi_6 - \tfrac{145}{36}\phi_8 + \phi_9 &= \tfrac{2}{3}e^{-2/3},\\
\phi_7 + \phi_8 - \tfrac{146}{36}\phi_9 + \phi_{10} &= \tfrac{1}{2} + \tfrac{5}{6}e^{-5/6},\\
\phi_9 - \tfrac{147}{36}\phi_{10} &= \tfrac{2}{3} + \tfrac{5}{6}e^{-5/6},
\end{aligned}
$$

where $\phi_1, \phi_2, \phi_3, \ldots, \phi_{10}$ are values of the function at the interior points.

The preceding system of equations written in matrix form gives

$$
\begin{bmatrix}
-\frac{135}{36} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -\frac{136}{36} & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -\frac{137}{36} & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & -\frac{138}{36} & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & -\frac{142}{36} & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & -\frac{143}{36} & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & -\frac{144}{36} & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & -\frac{145}{36} & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -\frac{146}{36} & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -\frac{147}{36}
\end{bmatrix}
\begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \\ \phi_7 \\ \phi_8 \\ \phi_9 \\ \phi_{10} \end{bmatrix}
=
\begin{bmatrix}
\frac{1}{3}e^{-1/3} \\
\frac{1}{2}e^{-1/2} \\
\frac{2}{3}e^{-2/3} \\
\frac{1}{6}+\frac{5}{6}e^{-5/6} \\
\frac{1}{2}e^{-1/2} \\
\frac{2}{3}e^{-2/3} \\
\frac{1}{3}+\frac{5}{6}e^{-5/6} \\
\frac{2}{3}e^{-2/3} \\
\frac{1}{2}+\frac{5}{6}e^{-5/6} \\
\frac{2}{3}+\frac{5}{6}e^{-5/6}
\end{bmatrix}
$$

We now rewrite the matrix on the preceding page in terms of block matrices, and this gives

$$\begin{bmatrix} A_{1,1} & A_{1,2} & 0 & 0 \\ A_{2,1} & A_{2,2} & A_{2,3} & 0 \\ 0 & A_{3,2} & A_{3,3} & A_{3,4} \\ 0 & 0 & A_{4,3} & A_{4,4} \end{bmatrix} \begin{bmatrix} \bar{\phi}_1 \\ \bar{\phi}_2 \\ \bar{\phi}_3 \\ \bar{\phi}_4 \end{bmatrix} = \begin{bmatrix} \bar{k}_1 \\ \bar{k}_2 \\ \bar{k}_3 \\ \bar{k}_4 \end{bmatrix},$$

where

$$\bar{\phi}_1 = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{bmatrix}, \quad \bar{\phi}_2 = \begin{bmatrix} \phi_5 \\ \phi_6 \\ \phi_7 \end{bmatrix}, \quad \bar{\phi}_3 = \begin{bmatrix} \phi_8 \\ \phi_9 \end{bmatrix}, \quad \bar{\phi}_4 = \begin{bmatrix} \phi_{10} \end{bmatrix},$$

$$\bar{k}_1 = \begin{bmatrix} \frac{e^{-1/3}}{3} \\ \frac{e^{-1/2}}{2} \\ \frac{2}{3} e^{-2/3} \\ \frac{1}{6} + \frac{5}{6} e^{-5/6} \end{bmatrix}, \quad \bar{k}_2 = \begin{bmatrix} \frac{e^{-1/2}}{2} \\ \frac{2}{3} e^{-2/3} \\ \frac{1}{3} + \frac{5}{6} e^{-5/6} \end{bmatrix}, \quad \bar{k}_3 = \begin{bmatrix} \frac{2}{3} e^{-2/3} \\ \frac{1}{2} + \frac{5}{6} e^{-5/6} \end{bmatrix}, \quad \bar{k}_4 = \begin{bmatrix} \frac{2}{3} + \frac{5}{6} e^{-5/6} \end{bmatrix},$$

$$A_{1,1} = \begin{bmatrix} -\frac{135}{36} & 1 & 0 & 0 \\ 1 & -\frac{136}{36} & 1 & 0 \\ 0 & 1 & -\frac{137}{36} & 1 \\ 0 & 0 & 1 & -\frac{138}{36} \end{bmatrix}, \quad A_{2,2} = \begin{bmatrix} -\frac{142}{36} & 1 & 0 \\ 1 & -\frac{143}{36} & 1 \\ 0 & 1 & -\frac{144}{36} \end{bmatrix},$$

$$A_{3,3} = \begin{bmatrix} -\frac{145}{36} & 1 \\ 1 & -\frac{146}{36} \end{bmatrix}, \quad A_{4,4} = \begin{bmatrix} -\frac{147}{36} \end{bmatrix}, \quad A_{2,1} = A_{1,2}^T = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$A_{3,2} = A_{2,3}^T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad A_{4,3} = A_{3,4}^T = \begin{bmatrix} 0 & 1 \end{bmatrix}.$$

Outline below is the general form for carrying the computation to completion after the matrices $A_{i,1}$ with $i,j=1,4$, and the vectors $\bar{k}_i$ with $i=1,4$ have been formed.

We first invert $A_{1,1}=L_{1,1}$ and multiply $L_{1,1}^{-1}$ by the vector $\bar{k}_1$. This gives the vector $\bar{\zeta}_1$ (equations 10), which we store because it is used later in the computation of the vectors $\bar{\phi}_i$. We now compute the product $L_{1,1}^{-1}A_{1,2}=U_{1,2}$, which is also stored because it is used in the computation of the vectors $\bar{\phi}_i$. We now multiply $U_{1,2}$ by $L_{2,1}=A_{2,1}$, and subtracting the resulting matrix from $A_{2,2}$, we have $L_{2,2}$, which we invert. The next computation necessary is the product $L_{2,1}\bar{\zeta}_1$ and the resulting vector is subtracted from $\bar{k}_2$. We multiply the vector $(\bar{k}_2-L_{2,1}\bar{\zeta}_1)$ just evaluated by $L_{2,2}^{-1}$, and this gives us the vector $\bar{\zeta}_2$. A repetition of the above operations will finally give us the vectors $\bar{\zeta}_1$, $\bar{\zeta}_2$, $\bar{\zeta}_3$, $\bar{\zeta}_4$, and the matrices $U_{1,2}$, $U_{2,3}$, $U_{3,4}$. This completes the first part of the computation. We are now in a position to compute the values of the function ϕ from equations 13.

From equation 13, we have $\bar{\phi}_4=\bar{\zeta}_4$. We form the product $U_{3,4}\bar{\phi}_4$, which we subtract from $\bar{\zeta}_3$ and obtain the vector $\bar{\phi}_3$. Continuing in this manner, we compute finally all the vectors $\bar{\phi}_1$, $\bar{\phi}_2$, $\bar{\phi}_3$, $\bar{\phi}_4$ which are the solution of our boundary value problem.

RESULTS AND DISCUSSION

In table 3 we have tabulated our results for the higher order corrections. The formula for $\Delta\delta_0$ has been given in the previous section.

The remaining two integrals are

$$\Delta^2\delta_0^{(2)} = -\frac{2}{k\sqrt{5}}\left(\int_0^R\int_0^{r_1}\Phi_0^{(0)}\frac{r_2^2}{r_1^3}\Phi_2^{(0)}\,dr_1\,dr_2+\int_R^\infty\int_0^{r_1}\Phi_0^{(0)}\frac{r_2^2}{r_1^3}\Phi_2^{(0)}\,dr_1\,dr_2\right), \tag{I-45}$$

$$\Delta^2\delta_1^{(1)} = -\frac{2}{k\sqrt{3}}\left(\int_0^R\int_0^{r_1}\Phi_0^{(0)}\frac{r_2^2}{r_1^3}\Phi_1^{(1)}\,dr_1\,dr_2+\int_R^\infty\int_0^{r_1}\Phi_0^{(0)}\frac{r_2^2}{r_1^3}\Phi_1^{(1)}\,dr_1\,dr_2\right). \tag{I-43}$$

TABLE 3.—*Resumé of Higher Order Results**

k*	$\Delta\delta_0$ 0-10	$\Delta\delta_0$ 10-∞	$\Delta^2\delta_0^{(2)}$ 0-10	$\Delta^2\delta_0^{(2)}$ 10-∞	$\Delta^2\delta_0^{(1)}$ 0-10	$\Delta^2\delta_0^{(1)}$ 10-∞
0	-0.20256	-0.3526	-0.012304	-0.00322	-0.001032	-0.00822
0.1	0.01895	0.01049	0.001168	0.000199	0.000369	0.000129
0.2	0.03136	0.00510	0.002016	0.000119	-0.000325	-0.000052
0.3	0.03530	0.00166	0.002418	0.000029	-0.00132	-0.000021
0.4	0.03373	0.00159	0.0024685	0.000026	-0.00228	+0.000019
0.5	0.03129	0.00188	0.002384	0.000041	-0.00140	0.0000049
0.6	0.02939	0.00131	0.002263	0.000028	-0.00141	-0.00000875
0.7	0.02766	0.00085	0.002145	0.000015	-0.00132	-0.0000003
0.8	0.02605	0.00096	0.002047	0.000019	-0.00130	-0.0000045

*For the purposes of the scattering length, the k = 0 entries are negative.

TABLE 4.—*The Convergence of the Nonadiabatic Series*

k	λ^0 δ_0	λ $\Delta\delta_0$	λ^2 $\Delta^2\delta_0^{(1)} + \Delta^2\delta_0^{(2)}$	δ
0	2.3482	-0.55516	-0.02477	1.7683
0.1	2.907728	0.02944	0.001865	2.9390
0.2	2.67942	0.03646	0.001758	2.7176
0.3	2.46158	0.03695	0.00111	2.4996
0.4	2.25800	0.03532	0.000235	2.2936
0.5	2.07102	0.03317	0.00104	2.1052
0.6	1.90189	0.03070	0.000872	1.9335
0.7	1.75070	0.02853	0.00084	1.7801
0.8	1.61666	0.02701	0.00076	1.6443

The significance of the break-up of the integrals at $R=10$ has also been explained. In this connection it can hardly be overstressed that for small k a significant contribution comes from the region $R \leq r_1 \leq \infty$.

In table 4 we have collected results to show the convergence of the terms multiplying successive powers of λ in the nonadabiatic series (see also, equation 43 in reference 2). The convergence appears to be even more rapid than an order of magnitude per power of λ. In second order this is due to a partial cancellation of the terms multiplying λ^2 a circumstance which may very well have an increasingly important effect for high powers of λ due to the increasing number of terms which enter. (It is also possible that it might have the opposite effect.) The final δ in each row is then the sum of the entires to its left to the accuracy that we believe it is meaningful. As a simple extrapolation of the convergence, we take this to mean an uncertainty of five units in the last figure given for the phase shifts. (The scattering length will be discussed below.)

It must be re-emphasized that this λ series is not identical to the more traditional series strictly in powers of l. The latter may be thought of as having been derived by truncating the original set of equations (see also, equation 4 of reference 2) after $l=\mathrm{L}$. That is,

$$\left[\Delta_{12} - l(l+1)\left(\frac{1}{r_1} + \frac{1}{r_2^2}\right) + \frac{2}{r_1} + \frac{2}{r_2} - M_{ll}\right]\Phi_l = \sum_{m=0}^{L} M_{lm}\Phi_m \qquad l = 0, 1, \cdots, \mathrm{L}\,.$$

Assuming we could solve each of these problems exactly, we would obtain a sequence of phase shifts $\delta_{(L)}$ which would approach the exact phase shift

$$\lim_{L\to\infty} \delta_{(L)} = \delta\,.$$

The $\delta_{(0)}$ is of course our δ_0. Beyond that Schwartz[11] has recently used his variational technique to solve for $\delta_{(1)}$.

There is, however, a somewhat more fundamental, albeit more idealized, sequence $\delta_{(L)}$ which can be defined. Assume we had the exact wave function $\Psi(r_1, r_2, \theta_{12})$. Then we could obtain the exact $\Phi_l(r_1 r_2)$ by suitably projecting $P_l(\cos\theta_{12})$ on Ψ. We could then obtain a sequence of $\delta_{(L)}$ from the basic relation

$$\sin(\delta_{(L)} - \delta_0) = -\frac{1}{k}\sum_l \frac{2}{\sqrt{2l+1}} \iint \Phi_0^{(0)} \frac{r_2^l}{r_1^{l+1}} \Phi_l \, dr_1 \, dr_2\,. \tag{I-13}$$

Clearly the second of these sequences of $\delta_{(L)}$ cannot be worse than the first (although the first obviously comes from a variational principle).

TABLE 5.—*Comparison of Schwartz's and Nonadiabatic Results*

k	Schwartz $\Delta\delta$	Nonadiabatic $\Delta\delta_0 + \Delta^2\delta_0^{(1)}$	Schwartz $\sum_{i=2}^{\infty} \Delta^{(i)}\delta$	Nonadiabatic $\Delta^2\delta_0^{(2)}$	Schwartz δ	Nonadiabatic δ
0	0.5670	0.5644	0.0126	0.01552	1.7686	1.7683
0.1		0.02986		0.001367	2.9388	2.9390
0.2	0.0362	0.03608	0.0015	0.002135	2.7171	2.7176
0.3		0.03561		0.002447	2.4996	2.4996
0.4	0.0340	0.03306	0.0018	0.002495	2.2938	2.2936
0.5		0.03178		0.002425	2.1046	2.1052
0.6	0.0302	0.03038	0.0008	0.002291	1.9329	1.9335
0.7		0.02721		0.002160	1.7797	1.7801
0.8	0.0250	0.02571	0.0017	0.002066	1.643	1.6444

In table 5 we have collated the results of Schwartz's and our calculations which bear on the latter sequence of $\delta_{(L)}$. The column marked $\Delta\delta$ is the difference $\delta_{(1)} - \delta_0$ in Schwartz's calculation.* The approximation here is the neglect of the back coupling of the higher Φ_l which distinguishes between the first and the second sequences of $\delta_{(L)}$. The analog of $\Delta\delta$ in our case is $\Delta^2\delta_0 + \Delta^2\delta_0^{(1)}$. Here the back-coupling is consistently taken into account, but we have only included two terms of a (presumably rapidly convergent) infinite series. In the column marked $\Delta^2\delta_0^{(2)}$ we have presented only the first term of the relevant infinite series. The corresponding column of Schwartz has been obtained by subtracting his final phase shifts from $\delta_{(1)}$. If his ansatz for the wave function were exact, we could conclude that this was the contribution of all remaining multipoles,

$$\sum_{i=2}^{\infty} \Delta^{(i)}\delta \,.$$

However, the ansatz for the complete wave function contains (presumably) about the same number of parameters as that used in obtaining $\delta_{(1)}$. Therefore, it is by no means clear that projecting out Φ_1 from his Ψ will give the same accuracy as his explicitly calculated $\delta_{(1)}$.

For these reasons it can hardly be expected that there would be equality between the corresponding entries in table 5. Nevertheless the rather wide deviation of the individual entries bespeaks of the possibility that the agreement to almost five significant figures in the final phase shifts may be somewhat coincidental. For the purposes of later discussion it should be noted that our quadrupole contribution is larger than all the remaining multipoles in Schwartz's calculation. (Thus, his results suggest a more rapid rate of convergence of the l-expansion than our own!)

Schwartz has also commented[11] on the relative angular momentum expansion in these types of problems. (The part of his scattering calculation which concerns the triplet phase shift has been given in table 5.) The bulk of his calculation is concerned with the second order energy (for the singlet spin state):

$$E_2 = \left(\sum_n + \int dn\right) \frac{\left|\left\langle 0 \left| \frac{2}{r_{12}} \right| n \right\rangle\right|^2}{E_0 - E_n} \,. \tag{15}$$

* Some time ago we requested Dr. Schwartz to use his Hylleraas variational approach to calculate δ_0 by omitting all terms depending on r_{12}. Dr. Schwartz kindly carried out these calculations which served as a check on our original zeroth order results (reference 2). The calculations were subsequently expanded to comprise the contents of tables 2 and 3 of reference 11.

Using the well known expansion of $2/r_{12}$ in Legendre polynomials,

$$\frac{2}{r_{12}} = 2\sum_{l=0}^{\infty} \frac{r_<^l}{r_>^{l+1}} P_l(\cos\theta_{12}) ,$$

Schwartz can put equation 15 into the form

$$E_2 = \sum_{l=0}^{\infty} E_2(l) ,$$

where the $E_2(l)$ can be well defined.[12] With each $E_2(l)$ there is associated a wave function $\Psi_1(l)$ which, aside from the angular dependence $P_l(\cos\theta_{12})$, is a function of the two radial variable r_1 and r_2. Reduced to its bare essentials, Schwartz's argument runs as follows: if we treat each l problem variationally with the usual type of smooth polynomial trial functions, then the $l=0$ and $l=1$ problems can be well approximated whereas the higher l problems become increasingly difficult. The reasons for the increasing difficulty of approximation by conventional means is due to the fact that the functions $\Psi_1(l)$ have discontinuities in their second derivates coming ultimately from different analytic forms of $r<^l/r<^{l+1}$ in the regions $r_1>r_2$ and $r_1<r_2$. The discontinuities correspond to the $\Psi_1(l)$ becoming more and more sharply peaked about the line $r_1=r_2$. On the basis that the bump itself provides the dominant contribution to the energy, Schwartz has derived the asymptotic formula for large l

$$E_2(l) \approx -\frac{45}{256}\frac{1}{l^4} . \tag{16}$$

This then, defines the convergence of this specific problem rather than any inaccurate calculations for $E_2(l)$ for $l>1$, which in general will tend to give the idea of a much more rapid convergence.*

Equation 16 applies to the specific problem of the second order energy in the singlet (space-symmetric) state; however, it is not unreasonable to assume l^{-n} characterizes the complete energy (or other physical property) in the l-expansion. Whether l^{-n} constitutes a rapidly convergent series depends on the type of problem at hand. In bound state problems where much greater experimental accuracy is in general available, it is necessary to be quite demanding in this regard. Even here, however, the convergence of equation 16 is not in principle uncompetitive with traditional techniques. Thus, if we associated the inclusion of an additional l component with the inclusion of an additional parameter in conventional expansions, in which the use of 100 (reference 5) and over 1000 parameters[13] has now been accomplished, a competitive 8 to 12 significant figure accuracy would be achieved. The fact that a (presently) conventional approach cannot be used in accomplishing this does not present an *a priori* objection to the rate of convergence of the l-expansion.

---

* Reference 11 must be read very carefully here; otherwise it may give the erroneous impression that

$$\sum_{l=2}^{4} E_2(l) \approx \frac{2}{3}\sum_{l=2}^{\infty} E_2(l) ,$$

which, if it were true, would be a more serious criticism of the convergence. What in fact is being asserted is that Schwartz's conventional calculation of

$$\sum_{l=2}^{4} E_2(l)$$

must be in error in such a way as to give a spurious rate of convergence. Nevertheless we can be quite sure that correctly

$$\sum_{l=2}^{4} E_2(l) \gtrsim (0.99)\sum_{l=2}^{\infty} E_2(l) .$$

In his discussion of the extension of the relative partial wave treatment to the scattering problem, we find that Schwartz has insufficiently stressed the different physics involved. First it is clear that because of the disparity in experimental accuracy not nearly the accuracy of a bound state problem is required to correlate theory with experiment.

In order further to discuss the scattering case, it is necessary to clarify the following point. In scattering calculations $r_1 r_2$ times the wave function has a nonvanishing component which must be included in order to make any kind of analysis. In discussing the aspects of $r_1 r_2 \Psi$ below, we shall always disregard this nonvanishing component.

$$\lim_{r_1 \to \infty} r_1 \, r_2 \Psi = A \sin(k r_1 + \delta) R_{1s}(r_2) \, .$$

The main difference between the bound state wave function (with any kind of forces) and scattering wave functions (involving Coulomb forces) is that the bound-state wave function vanishes exponentially in all asymptotic regions of configuration space whereas the scattering wave function does not. It has been one of the primary points of the nonadiabatic theory in the decomposition of $r_1 r_2 \Psi$ in terms of $P_l(\cos\theta_{12})$, equation I-3, that the associated Φ_l can be shown to have slowly vanishing adiabatic forms

$$\lim_{r_1 \to \infty} \Phi_l \approx - \frac{2}{\sqrt{2l+1}} \frac{\sin(k r_1 + \delta)}{r_1^{l+1}} e^{-r_2} \left(\frac{r_2^{l+2}}{l+1} + \frac{r_2^{l+1}}{l} \right) .$$

It has further been derived as one of the main results of this theory that the scattering length due to the dipole term will be diminished by an amount[7]

$$a = a(R) - \frac{9}{2} \left(\frac{1}{R} - \frac{a + a_0}{2R^2} + \cdots \right) ,$$

where $a(R)$ comes from a wave function which is more sharply cut off and hence more characteristic of a problem in a finite "sphere" of radius R. This prediction (reference 7 and A. Temkin personal communication to C. Schwartz) was tacitly confirmed by the calculation of Schwartz[4] in which, when the variational counterpart of Φ_1 was included, his scattering length was reduced by over 5 percent. Thus, whereas the nonadiabatic theory incorporates both short-range correlations (via the convergent expansion in l) and long-range effects naturally, the Hylleraas type wave function by itself cannot practically deal with the latter. [Tables 2 and 3 of Schwartz's paper[11] include Φ_1 in the calculations of the $k=0$ entries for $\lambda(s+p)$ as well as λ (complete).]

That a Hylleraas type wave function does not naturally describe the long-range correlations can be further brought home by reference to the paper of Ohmura and Ohmura.[14] In their clever deduction of the singlet scattering length, these authors required the coefficient $C(\infty)$ in the adiabatic form

$$\lim_{r_1 \to \infty} \Psi = C(\infty) \frac{e^{-\gamma r_1}}{r_1} \frac{R_{1s}(r_2)}{r_2} \tag{17}$$

of the H^- wave function. Here γ is the square root of the electron affinity and being small it makes the term simulate the nonvanishing term in a scattering calculation. This form is not the analytic form of the Pekeris or Hylleraas wave functions both of which have the exponential dependence

$$\exp\left[- \frac{|E|^{1/2}}{2} (r_1 + r_2) \right] ,$$

where E is the total energy of the H^- ion. In order to evaluate $C(\infty)$, Ohmura and Ohmura used the 161 and 203 parameter Pekeris wave functions.[5] They concluded that, whereas the wave function reproduces

the adiabatic form (equation 17) quite accurately in the region $r_1 = 10$ to 12 ($r_2 = 0$), deviations in the variational approximations for $r_1 \geq 12$ were quite noticeable!

We shall now show that the inability of Schwartz's zero energy wave function to describe the adiabatic part of the quadrupole term can explain the difference between his triplet scattering length 1.7686, and our own 1.7683. To repeat, at zero energy Schwartz's Ψ is made to contain $\Phi_1^{(adaib)}$ but not $\Phi_2^{(adaib)}$. Using the same type analysis that we made on the dipole effect, we can write the long-range contribution of the quadrupole term as

$$a(R) - a = \lim_{k \to 0} \frac{2}{k^2 \sqrt{5}} \int_R^\infty \int_0^{r_1} \Phi_0^{(0)} \frac{r_2^2}{r_1^3} \Phi_2 \, dr_1 \, dr_2 ,$$

$$\approx \left[\frac{15}{k^2} \int_R^\infty \frac{\sin(kr + \delta_0) \sin(kr + \delta) \, dr}{r^6} \right]_{k \to 0} ,$$

$$\approx 15 \left[\frac{1}{3R^3} - \frac{1}{4} \frac{(a + a_0)}{R^4} + \cdots \right] .$$

If we associate Schwartz's value with $a(R)$, a very reasonable choice of $R (R \approx 25)$ will reproduce our own value for a. Although from the experimental point of view the difference between the two numbers is completely negligible, it is worth noting that the new scattering length is outside of Schwartz's limit of error.[2] We are inclined to think that the unaccounted for higher multipole may contribute an additional couple of units in the last place (see also, the $k = 0$ row of table 4). We would extrapolate the triplet scattering length to be:

$$a_t = 1.7675 \pm .0005 .$$

A question remains at non-zero energies as to the effect of the slowly vanishing multipoles. The answer obviously depends on the accuracy in question. Schwartz claims[4] that the effects are "washed out" to his accuracy whereas our own calculation suggests that particularly the dipole contribution is not. It may very well be that our inclusion of these effects plus Schwartz's inclusion of the higher multipoles balance each other out, and that both calculations give lower bounds for the phase shifts.

ACKNOWLEDGMENTS

The authors would like to thank Mrs. Dorothy Hoover who extended the zeroth program to double precision. The authors are particularly indebted to Mr. William F. Cahill, for his invaluable and generous help and advice on all aspects of the numerical solution of the partial differential equations.

REFERENCES

1. Temkin, A., "Nonadiabatic Theory of the Scattering of Electrons from Hydrogen," *Phys. Rev. Letters* **4**(11): 566–568, June 1, 1960.
2. Temkin, A., "Nonadiabatic Theory of Electron-Hydrogen Scattering," NASA Technical Note #ND-1343, July 1962. Also published in *Phys. Rev.* **126**(1): 130–142, April 1, 1962.
3. Clarke, P. G., and Schey, H. M., "Elastic Scattering of Low-Energy Electrons by Atomic Hydrogen," *Phys. Rev.* **126**(1): 147–162, April 1, 1962.
4. Schwartz, C., "Electron Scattering from Hydrogen," *Phys. Rev.* **124**(5): 1468–1471, December 1, 1961.
5. Pekeris, C. L., "Ground State of Two-Electron Atoms," *Phys. Rev.* **112**(5): 1649–1658, December 1, 1958.
6. Schwartz, C., "Variational Calculations of Scattering," *Annals of Physics* **16**(1): 36–50, October 1961.
7. Temkin, A., "Polarization and the Triplet Electron-Hydrogen Scattering Length," *Phys. Rev. Letters* **6**(7): 354–355, April 1, 1961.

8. TEMKIN, A., and HOOVER, D. E., "Nonseparable Theory of Electron-Hydrogen Scattering," in: *Methods in Computational Physics*, ed. by B. Alder, S. Fernbach, and M. Rotenberg, Vol. 1, New York: Academic Press. (In Press.)
9. FADDEEVA, V. N., "Computational Methods of Linear Algebra," Translated by C. D. Benster, New York: Dover, 1959, p. 118.
10. HOUSEHOLDER, A. S., "The Theory of Matrices in Numerical Analysis," in: *University of Michigan Engineering Summer Conferences, 1962: Numerical Analysis.*
11. SCHWARTZ, C., "Importance of Angular Correlations Between Atomic Electrons," *Phys. Rev.* **126**(3): 1015–1019, May 1, 1962.
12. SCHWARTZ, C., "Estimating Convergence Rates of Variational Calculations," in: *Methods in Computational Physics*, ed. by B. Alder, S. Fernbach, and M. Rotenberg, Vol. 1, New York: Academic Press. (In Press.)
13. PEKERIS, C. L., "$1\,{}^1S$ and $2\,{}^3S$ States of Helium," *Phys. Rev.* **115**(5): 1216–1221, September 1, 1959.
14. OHMURA, T., and OHMURA, H., "Electron-Hydrogen Scattering at Low Energies," *Phys. Rev.* **118**(1): 154–157, April 1, 1960.

ACKNOWLEDGMENT OF NON-NASA PUBLICATIONS

The following is a listing of the non-NASA sources from which the various articles reprinted in this book have been taken.

American Institute of Aeronautics and Astronautics Journal

STRUCTURE OF PLANETARY ATMOSPHERES, *by* S. I. RASOOL

Anglo-American Aeronautical Conference, 9th, Cambridge and Montreal, October 1963

SCIENTIFIC RESULTS OF OSO–I, *by* J. C. LINDSAY

Annales d'Astrophysique

THE CHARACTERISTICS OF DISTANT COMETS, *by* B. DONN

Annals of Physics

ON THE MOTION OF CHARGED PARTICLES IN THE GEOMAGNETIC FIELD, *by* E. C. RAY

THEOREMS CONCERNING THE ALLOWED CONE OF COSMIC RADIATION, *by* E. C. RAY

ON COSMIC RAY CUTOFFS, *by* H. H. SAUER and E. C. RAY

Astronautics and Aerospace Engineering

RESEARCH WITH TIROS RADIATION MEASUREMENTS, *by* W. NORDBERG

Astronomical Journal

PRELIMINARY STUDY OF PERIODIC ORBITS OF INTEREST FOR MOON PROBES. II, *by* S.-S. HUANG and C. WADE, Jr.

Astrophysical Journal

NUCLEAR COMPOSITION AND RIGIDITY SPECTRA OF SOLAR COSMIC RAYS, *by* S. BISWAS and C. E. FICHTEL

NEUTRINO PROCESSES AND RED GIANTS, *by* H.-Y. CHIU

AN INTERPRETATION OF BETA LYRAE, *by* S.-S. HUANG

MODES OF MASS EJECTION BY BINARY STARS AND THE EFFECT ON THEIR ORBITAL PERIODS, *by* S.-S. HUANG

TRANSFER OF ANGULAR MOMENTUM BETWEEN EJECTED PARTICLES AND THE BINARY SYSTEM, *by* S.-S. HUANG

EVOLUTION OF O STARS. I. HYDROGEN-BURNING, *by* R. STOTHERS

NEUTRINO EMISSION, MASS LOSS, AND THE FREQUENCY OF SUPERNOVAE, *by* R. STOTHERS

Dynamics of Satellites: IUTAM Symposium, Paris, 1962

SATELLITE ORBIT ANALYSES FOR GEODETIC PURPOSES, *by* W. M. KAULA

ON LONG RANGE EFFECTS IN THE MOTION OF ARTIFICIAL SATELLITES, *by* P. MUSEN

Icarus

FORMATION OF THE SOLAR NEBULA, *by* A. G. W. CAMERON

THE ORIGIN OF THE ATMOSPHERES OF VENUS AND THE EARTH, *by* A. G. W. CAMERON

THE ORIGIN AND STRUCTURE OF ICY COMETARY NUCLEI, *by* B. DONN

THE EARLY EVOLUTION OF THE SUN, *by* D. EZER and A. G. W. CAMERON
THE RELATION OF TEKTITES TO LUNAR IGNEOUS ACTIVITY, *by* P. D. LOWMAN, Jr.

Nature

INTERPLANETARY MAGNETIC FIELDS AS A CAUSE OF COMET TAILS, *by* D. B. BEARD and M. P. NAKADA
ENHANCEMENT OF LIGHT OUTPUT FROM A SUPERNOVA, *by* S. A. COLGATE and A. G. W. CAMERON

Journal of the Atmospheric Sciences

EXPERIMENTAL CONFIRMATION FROM THE TIROS VII METEOROLOGICAL SATELLITE OF THE THEORETICALLY CALCULATED RADIANCE OF THE EARTH WITHIN THE 15-MICRON BAND OF CARBON DIOXIDE, *by* W. R. BANDEEN, B. J. CONRATH, and R. A. HANEL
PLASMA DIFFUSION IN THE IONOSPHERE, *by* S. CHANDRA
THE INFRARED HORIZON OF THE PLANET EARTH, *by* R. A. HANEL, W. R. BANDEEN, and B. J. CONRATH
INVESTIGATION OF PROBLEMS IN THERMAL CONVECTION, *by* J. R. HERRING
EVIDENCE FOR THE EFFECT OF CORPUSCULAR RADIATION ON THE IONOSPHERE, *by* F. MARIANI

Journal of Chemical Physics

THE STRUCTURE OF ZINC OXIDE NUCLEI, *by* G. W. SEARS, R. POWELL, and B. DONN
RELATIVE PARTIAL WAVE THEORY OF DIATOMIC MOLECULES, *by* A. TEMKIN

Journal of Geophysical Research

SIMULTANEOUS ROCKET AND SATELLITE MEASUREMENTS OF THE TOPSIDE IONOSPHERE, *by* S. J. BAUER
SOME IMPLICATIONS OF A DIRECT MEASUREMENT OF THE HYDROGEN AND HELIUM ION DISTRIBUTION IN THE UPPER ATMOSPHERE, *by* S. J. BAUER
POLARIZATION OF SYNCHROTRON RADIATION AND THE QUASI-TRANSVERSE PROPAGATION REGION, *by* L. J. BLUMLE
IONOSPHERE ELECTRON TEMPERATURE MEASUREMENTS AND THEIR IMPLICATIONS, *by* L. H. BRACE, N. W. SPENCER, and G. R. CARIGNAN
AUTOMATIC MAPPING OF THE GEOMAGNETIC FIELD, *by* J. C. CAIN and J. R. NEILON
NEGATIVE CORRELATION BETWEEN POLAR-CAP VISUAL AURORA AND MAGNETIC ACTIVITY, *by* T. N. DAVIS
TWENTY-FOUR-HOUR OBSERVATIONS OF AURORA AT THE SOUTHERN AURORAL ZONE, *by* T. N. DAVIS and R. N. DEWITT
ANALYSIS OF TOPSIDE SOUNDER RECORDS, *by* R. J. FITZENREITER and L. J. BLUMLE
OBSERVATIONS ON THE SOLAR PARTICLE EVENTS OF JULY 1961, *by* D. E. GUSS and C. J. WADDINGTON
RELATION BETWEEN THEORETICAL AND OBSERVATIONAL MODELS OF THE UPPER ATMOSPHERE, *by* I. HARRIS and W. PRIESTER
THE ROLE OF PLASMA INSTABILITIES IN THE ORIGIN OF SOLAR FLARES, *by* R. K. JAGGI
HYDROGEN DIFFUSION IN THE THERMOSPHERE BY MONTE CARLO METHOD, *by* S. O. KASTNER
ELASTIC MODELS OF THE MANTLE CORRESPONDING TO VARIATIONS IN THE EXTERNAL GRAVITY FIELD, *by* W. M. KAULA
IMPROVED GEODETIC RESULTS FROM CAMERA OBSERVATIONS OF SATELLITES, *by* W. M. KAULA

TIDAL DISSIPATION IN THE MOON, *by* W. M. KAULA

ELECTRONS FROM BOMB NEUTRON DECAY, *by* J. KILLEEN, W. N. HESS, and R. E. LINGENFELTER

IMPULSIVE INJECTION OF ELECTRONS INTO THE EARTH'S INNER TRAPPING REGION FROM SOLAR COSMIC-RAY EVENTS, *by* M. LIWSHITZ and A. M. LENCHEK

AURORAL DISSOCIATION OF MOLECULAR OXYGEN IN THE POLAR MESOSPHERE, *by* K. MAEDA

ON THE ZENITHAL DISTRIBUTION OF EXTREMELY HIGH ENERGY COSMIC RAY MUONS IN THE ATMOSPHERE, *by* K. MAEDA

PITCH ANGLE DISTRIBUTION OF THE PHOTOELECTRONS AND ORIGIN OF THE GEOMAGNETIC ANOMALY IN THE F2 LAYER, *by* F. MARIANI

THE SHAPE OF THE GEOMAGNETIC-FIELD SOLAR-WIND BOUNDARY, *by* G. D. MEAD and D. B. BEARD

ON DETERMINING THE SECULAR AND CRITICAL EFFECTS IN THE MOTION OF SATELLITES BY MEANS OF A NONSINGULAR SET OF VECTORIAL ELEMENTS, *by* P. MUSEN

ON THE GENERAL PLANETARY PERTURBATIONS IN RECTANGULAR COORDINATES, *by* P. MUSEN and L. CARPENTER

HIGH-ENERGY ELECTRONS IN THE RADIATION BELT, *by* M. P. NAKADA

SYNCHROTRON RADIATION CALCULATIONS FOR THE ARTIFICIAL RADIATION BELT, *by* M. P. NAKADA

VARIATIONS OF THE PROTON ENERGY SPECTRUM WITH POSITION IN THE INNER VAN ALLEN BELT, *by* J. E. NAUGLE and D. A. KNIFFEN

SOLAR PROTON SPECTRUMS IN THE EVENTS OF NOVEMBER 12 and 15, 1960, *by* K. W. OGILVIE and D. A. BRYANT

A MODEL OF THE QUIET IONOSPHERE, *by* J. C. SEDDON

OSCILLATION OF THE GEOMAGNETIC FIELD LINES AND ASSOCIATED MAGNETIC PERTURBATIONS AT CONJUGATE POINTS, *by* M. SUGIURA and C. R. WILSON

DIRECT MEASUREMENTS OF HELIUM AND HYDROGEN ION CONCENTRATION AND TOTAL ION DENSITY TO AN ALTITUDE OF 940 KILOMETERS, *by* H. A. TAYLOR, Jr., L. H. BRACE, H. C. BRINTON, and C. R. SMITH

TRAPPED ELECTRON TIME HISTORIES FOR $L=1.18$ to $L=1.30$, *by* J. A. WELCH, Jr., R. L. KAUFMANN, and W. N. HESS

DISCUSSION OF OUR EARLIER PAPER 'HYDROMAGNETIC INTERPRETATION OF SUDDEN COMMENCEMENTS OF MAGNETIC STORMS,' *by* C. R. WILSON and M. SUGIURA

HYDROMAGNETIC WAVES GENERATED BY THE JULY 9, 1962, NUCLEAR WEAPONS TEST AS OBSERVED AT COLLEGE, ALASKA, *by* C. R. WILSON and M. SUGIURA

Methods in Computational Physics

NONSEPARABLE THEORY OF ELECTRON-HYDROGEN SCATTERING, *by* A. TEMKIN and D. E. HOOVER

The Physical Review

COSMIC-RAY HYDROGEN AND HELIUM NUCLEI DURING A SOLAR QUIET TIME IN JULY 1961, *by* C. E. FICHTEL, D. E. GUSS, G. R. STEVENSON, and C. J. WADDINGTON

ELECTRON IMPACT 2s and 2p EXCITATIONS OF ATOMIC HYDROGEN, *by* K. OMIDVAR

Physical Review Letters

NEW EVIDENCE FOR LONG-LIVED SOLAR STREAMS IN INTERPLANETARY SPACE, *by* D. A. BRYANT, T. L. CLINE, U. D. DESAI, and F. B. MCDONALD

VERY HIGH-ENERGY COSMIC RAYS, *by* C. E. FICHTEL

Physical Society (London) Proceedings

REARRANGEMENT COLLISIONS III. ELECTRON TRANSFER FROM ATOMS TO FAST PROTONS, *by* M. H. MITTLEMAN

Radio Science Journal of Research

CONSTITUTION OF THE ATMOSPHERE AT MAGNETOSPHERIC LEVELS, *by* S. J. BAUER

Reviews of Geophysics

DETERMINATION OF THE EARTH'S GRAVITATIONAL FIELD, *by* W. M. KAULA

Science

INTERPLANETARY DUST, *by* W. M. ALEXANDER

PLANETS AND COMETS: ROLE OF CRYSTAL GROWTH IN THEIR FORMATION, *by* B. DONN and G. W. SEARS

STRESS DIFFERENCES AND THE REFERENCE ELLIPSOID, *by* J. A. O'KEEFE and W. M. KAULA

ORIGIN OF TEKTITES, *by* J. A. O'KEEFE and B. E. SHUTE

Sky and Telescope

JEANS' CRITERION OF GRAVITATIONAL INSTABILITY, *by* S.-S. HUANG

RECENT ADVANCES CONCERNING THE EARLY HISTORY OF THE SOLAR SYSTEM, *by* S.-S. HUANG

Space Science Reviews

THE WORLD MAGNETIC SURVEY, *by* J. P. HEPPNER

Transactions American Geophysical Union

ROCKET AND SATELLITE INVESTIGATIONS OF THE IONOSPHERE, *by* R. E. BOURDEAU

GEOMAGNETIC DISTURBANCES, *by* M. SUGIURA

AUTHOR INDEX

☆ U.S. GOVERNMENT PRINTING OFFICE : 1965 O—759-522